University of Zurich[UZH]

Department of Informatics

Computational frameworks to increase effectiveness and efficiency of data collection in Life Cycle Assessment

Dissertation submitted to the Faculty of Business,
Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wissenschaften, Dr. sc.
(corresponds to Doctor of Science, PhD)

presented by
Jürgen Reinhard
from Germany

approved in April 2017

at the request of
Prof. Dr. Lorenz M. Hilty
Prof. Göran Finnveden, Ph.D.
Prof. Dr. Burkhard Stiller

University of Zurich[UZH]

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, April 5, 2017

Chair of the Doctoral Board: Prof. Dr. Michael Böhlen

"I now walk into the wild"

-- Christopher McCandless --

Acknowledgements

First and foremost, I should like to thank my supervisor Prof. Dr. Lorenz M. Hilty. Researching and writing this dissertation would have been a far less pleasant task without his constant support, his sharp-witted feedback, and his granting me the freedom to explore. Lorenz, it has been an honor to be your first Ph.D. student at the University of Zurich.

I gratefully acknowledge my external and internal co-examiners, Prof. Dr. Göran Finnveden and Prof. Dr. Burkhard Stiller, for their willingness to carefully review my dissertation, and Prof. Dr. Michael Böhlen for his willingness to act as Chair of the Doctoral Board. I am honored to have such an outstanding committee assess my thesis.

My gratitude goes to Empa—the Swiss Federal Laboratories for Materials Science and Technology—for providing me with excellent infrastructure and generous funding for this dissertation. I especially appreciate the support given by the heads of the Technology and Society Lab, Heinz Böni and Dr. Patrick Wäger.

I am grateful to Dr. Rainer Zah for providing the flexible working conditions that allowed me to simultaneously work on my doctoral degree and hold a position at Quantis. I am also very grateful for his co-authorship and for his unwavering support throughout the years.

I wish to thank Dr. Gregor Wernet for his co-authorship and for appointing me to ecoinvent's group of LCI experts. My inspiring discussions with this group over the last six years have provided me with important technical insights that were necessary for the writing of this dissertation. Furthermore, I would like to thank Dr. Christopher Mutel for his co-authorship, stimulating input, and critical insights.

I would like to thank my co-workers and friends Andrea, Andi, Simon, Mireille, Emilia, Tereza, Dominic, and Hans-Jörg for their companionship throughout the years. My thanks go to Angela Adams for proofreading and Barbara Jost for her support in organizing my graduation.

I am most thankful to my parents and sisters for their love, continuous support, and encouragement over the years.

Finally, I wish to express my deepest gratitude to Melanie and to my children Jaron, Ruby, and Miron for sharing this dream with me, and bearing with me during its realization. It is to them that I lovingly dedicate this dissertation.

Abstract

Life Cycle Assessment (LCA) is perhaps the leading technique for assessing the environmental impact of products over their entire life cycle, and currently plays a decisive role in the field of environmental management. The technique is highly data-intensive and would be practically impossible without using the background unit process datasets provided by Life Cycle Inventory (LCI) databases. A typical product life cycle covers thousands of unit processes, each of which needs to be described with exchange flow data. Although the increased availability of LCI databases has substantially decreased the amount of work involved in conducting an LCA, the high costs associated with data collection still impose substantial limitations on the quality of LCA. These limitations either take the form of uncertain or incomplete background data, which must be gathered and maintained at great effort, or an excessively generic representation of highly context-dependent activities, for example, agricultural work.

The overall goal of this dissertation is to increase the efficiency and effectiveness of data collection in LCA by means of two novel and complementary solutions based on information systems (IS). The first solution is a statistical prioritization approach that directs LCI database improvement efforts toward datasets of key importance in terms of their potential influence on overall database quality. This approach aims to increase the effectiveness of data collection for LCI databases. The second solution is a computational framework that facilitates the automated generation of regionalized cultivation datasets on the basis of publicly available spatial (raster) data. This approach aims to increase the efficiency and quality of unit process dataset generation in the important domain of agricultural LCAs.

We examine the first, statistical prioritization approach in detail in two research articles. Article I (Chapter 2) presents a computational framework for prioritizing LCI database improvements. We demonstrate, then evaluate, a method for database-wide contribution analysis (CA) and corresponding summary measures that facilitate the identification of key processes, that is to say, unit processes with consistently large relative contributions throughout all product systems in the database. We show that prioritizing the improvement efforts is very useful because a tiny, robust nucleus of unit processes proves to be consistently important across all product systems and many Life Cycle Impact Assessment (LCIA) indicators. Focusing research efforts on these processes makes it possible to improve the LCI database effectively.

Article II (Chapter 3) applies the same statistical prioritization framework to the ecoinvent databases in a comprehensive case study. We identify the most important unit processes according to a set of 19 selected LCIA indicators using a newly developed ranking algorithm. Our study shows that a relatively large proportion of the overall database quality is dependent on a small set of key processes. Overall, 300 (out of 11,000) datasets cause 60% of the environmental impacts across all LCIA indicators, while just three datasets cause 11% of all climate change impacts. We present a ranking of key processes that adds a new perspective to database improvements, in that it makes it possible to allocate resources according to the structural dependencies in the data.

We examine the second, regionalization approach in another research article. Article III (Chapter 4) presents a computational framework that allows the automated, site-specific (regionalized) generation and assessment of cradle-to-gate agricultural unit process datasets. The framework facilitates the transformation of publicly available spatial (raster) data into comprehensive unit process datasets using default data from Version 3.2 of the ecoinvent database and the emission models from the World Food Life Cycle Database (WFLDB) guidelines. To illustrate the application of our framework, we describe its key features and present a case study on rapeseed production in Germany. Our study shows that automatically generating regionalized cultivation datasets harbors great potential for improving the accuracy of agricultural unit process datasets in LCA applications. With 580,000 datasets, the case study presented is likely the most comprehensive cradle-to-gate LCA on the climate change and eutrophication impacts of rapeseed cultivation in Germany.

Our research demonstrates that both prioritization and regionalization are valid, useful approaches to improving the data foundation of LCA-based decision-making. The prioritization framework makes it easier to align data collection efforts in LCI databases with those datasets that are the most important in terms of their overall influence on the quality of the database. Focusing research efforts on these processes makes it possible to effectively improve the datasets that play a dominant role in nearly every LCA application. Our research provides valuable new design knowledge in the form of operational principles that can be applied and adapted in other, as yet unstudied fields. Our framework for regionalized LCI modeling makes it easier to automatically generate high-resolution agricultural process datasets for all major crops and all regions around the world. The framework improves the geographical representativeness and reproducibility of agricultural datasets and offers new possibilities for their aggregation and analysis. Finally, the design knowledge that we develop provides a starting point for enhancing the utility of LCA, particularly in the context of other environmental system analysis tools.

Zusammenfassung

Die Ökobilanzierung (engl. Life Cycle Assessment, LCA) ist vermutlich die bedeutendste Methode zur Beurteilung der Umweltauswirkung von Produkten über ihren gesamten Lebensweg und spielt eine entscheidende Rolle im Umweltmanagement. LCA ist sehr datenintensiv und wäre ohne die Nutzung von modularen Datensätzen aus Inventardatenbanken—sogenannten Hintergrunddaten—praktisch unmöglich. Ein typischer Produktlebensweg beinhaltet tausende von Datensätze und für jeden einzelnen davon muss der detaillierte Stoffwechsel—Material-und Energieströme—beschrieben werden. Die zunehmende Verfügbarkeit von Inventardatenbanken hat den Zeit- und Kostenaufwand für die Durchführung von Ökobilanzen wesentlich reduziert. Die hohen Kosten der Datensammlung führen aber immer noch zu starken Qualitätseinbussen bei Ökobilanzstudien. Einerseits in der Form von unsicheren oder unvollständigen Hintergrunddaten, die mit besonders grossem Aufwand gesammelt und gepflegt werden müssen. Andererseits durch die zu generische Abbildung von hochgradig kontextabhängigen Aktivitäten, die—eben auch aufgrund der hohen Kosten in der Datensammlung—auf Durchschnitte anstatt auf Genauigkeit fokussieren müssen.

Das Ziel dieser Doktorarbeit ist die Steigerung der Effizienz und der Effektivität der Datensammlung in der Ökobilanzierung durch die Entwicklung von zwei neuen und komplementären IT-Lösungen. Einerseits wird ein statistischer Priorisierungsansatz entwickelt, der Verbesserungsbestrebungen in Inventardatenbanken auf die Kerndatensätze lenkt, die im Hinblick auf ihren potentiellen Einfluss auf die gesamte Datenqualität von entscheidender Bedeutung sind. Dieser Ansatz zielt darauf ab, die Effektivität der Datensammlung in Inventardatenbanken zu verbessern. Andererseits wird ein computergestütztes Verfahren entwickelt, das die automatische Generierung von regionalisierten Anbaudatensätzen auf Basis von öffentlich verfügbaren, räumlichen Rasterdatensätzen ermöglicht. Dieser Ansatz zielt darauf ab, die Effizienz und die Qualität der Ökobilanzierung im Agrarbereich zu steigern.

Wir behandeln den statistischen Priorisierungsansatz in der Form von zwei Forschungsartikeln. Artikel I (*Kapitel 2*) präsentiert unseren Ansatz zur Priorisierung von Datensatzverbesserungen. Wir demonstrieren und evaluieren eine Methode zur datenbankweiten Anwendung einer relativen Beitragsanalyse und zugehörigen Indikatoren zur Identifikation von Kerndatensätzen. Diese Datensätze zeichnen sich durch einen relativ grossen Beitrag zur Gesamtumweltauswirkung über zahlreiche Produktsysteme in der Inventardatenbank aus. Unsere Studie zeigt, dass die

Priorisierung von Datenbankverbesserungen grossen Nutzen hat. Ein robuster und kleiner Kern an Datensätzen ist, trotz Anwendung zahlreicher Wirkungsabschätzungsindikatoren, von konsistent grosser Bedeutung in allen Produktsysteme in der Datenbank. Die Ausrichtung zukünftiger Forschungsbemühungen auf die Verbesserung dieser Datensätze ermöglicht die effektive Verbesserung der gesamten Inventardatenbank.

Artikel II (Kapitel 3) wendet das entwickelte Priorisierungsverfahren in einer umfassenden Fallstudie auf die ecoinvent Datenbank an. Wir identifizieren und präsentieren die wichtigsten Kerndatensätze unter Berücksichtigung von 19 ausgewählten Wirkungsabschätzungsindikatoren und unter Anwendung eines neu entwickelten Ordnungsalgorithmus. Unsere Studie zeigt, dass ein relativer grosser Anteil der gesamten Datenbankqualität von der Qualität einer kleinen Anzahl von Kerndatensätzen abhängig ist. Im Grossen und Ganzen verursachen ca. 300 (von insgesamt 11'000) Datensätze ca. 60% der gesamten Umweltauswirkungen über alle berücksichtigten Wirkungsabschätzungsindikatoren, während 11% der gesamten Klimawirkung von nur drei Datensätzen verursacht wird. Wir präsentieren eine Rangliste von Kerndatensätzen, die eine neue Perspektive auf Datenbankverbesserungen zulässt, eben weil die Zuordnung der verfügbaren Arbeitsressourcen gemäss der strukturellen Bedeutung der Datensätze vorgenommen werden kann.

Wir behandeln den Ansatz zur automatischen Regionalisierung von Agrardatensätzen in Form eines Forschungsartikels. Artikel III (Kapitel 4) präsentiert ein Verfahren, das die automatische, standortspezifische (regionalisierte) Generierung und Umweltbeurteilung von Agrardatensätzen ermöglicht. Das Verfahren kann, unter Verwendung von Standarddaten der ecoinvent Datenbank und den Emissionsmodellen der World Food Life Cycle Database Guidelines, öffentlich verfügbare Rasterdaten in vollständige und regionalisierte Agrardatensätze konvertieren. Wir beschreiben die Hauptmerkmale des Ansatzes und präsentieren eine Fallstudie zum Rapsanbau in Deutschland, um dessen Anwendung in der Praxis zu demonstrieren. Unsere Studie zeigt, dass die computergestützte Generierung von regionalisierten Agrardatensätzen grosses Potential zur Verbesserung der Genauigkeit von Agrardatensätzen in Ökobilanzstudien beherbergt. Mit 580'000 Datensätzen ist die Fallstudie vermutlich die umfangreichste Ökobilanz zu den potentiellen Klima- und Eutrophierungsauswirkungen des Rapsanbaus in Deutschland.

Unsere Forschung hat gezeigt, dass Beides, Priorisierung und Regionalisierung wirkungsvolle und nützliche Ansätze zur Verbesserung der Datengrundlage der ökobilanzbasierten Entscheidungsfindung darstellen. Der Priorisierungsansatz ermöglicht die Ausrichtung von

Datensammlungen auf die Datensätze, die den grössten Einfluss auf die gesamte Datenqualität haben. Die Konzentration der Forschungsbestrebungen auf diese Prozesse ermöglicht die effektive Verbesserung der Kerndatensätze, die in praktisch jeder Ökobilanzanwendung eine dominante Rolle spielen. Unsere Forschung liefert neues und nutzbringendes Gestaltungswissen in Form von operativen Prinzipien, das auf andere Bereiche übertragen oder an andere Fragestellungen angepasst werden kann. Unser Ansatz zur regionalisierten Inventarmodellierung ermöglicht die automatische Generierung von hoch aufgelösten Agrarprozessen für alle wichtigen Pflanzen und für alle Gebiete der Welt. Der Ansatz verbessert den räumlichen Repräsentationsgrad und die Vergleichbarkeit von Agrardatensätzen und eröffnet neue Möglichkeiten für deren Aggregation und Analyse. Das entwickelte Gestaltungswissen liefert wichtige Ansatzpunkte zur Verbesserung des Nutzwertes der Ökobilanzierung, insbesondere im Rahmen von anderen Umweltsystemanalyseanwendungen.

TABLE OF CONTENTS

PART I – SYNOPSIS

1 SYNOPSIS

1.1 Introduction

1.1.1 The complexity of product systems

In today's economic systems, the provision, use, and disposal of almost any product involves complex globalized networks, known as "product systems" (Hellweg and Milà i Canals, 2014). Even for a relatively simply product such as a pencil, "not a single person on the face of this earth knows how to make [it]" (Read L., 1958, p. 4). This is simply because the product system involves countless interlinked human activities that are distributed across space and time. We can get a glimpse of this complexity by envisioning just a few of the activities involved in the product system:

"(…) the saws and trucks and rope and the countless other gear used in harvesting and carting the cedar logs to the railroad siding (…) the mining of ore, the making of steel and its refinement into saws, axes, motors; the growing of hemp and bringing it through all the stages to heavy and strong rope; the logging camps with their beds and mess halls, the cookery and the raising of all the foods (…) the heat, the light and power, the belts, motors, and all the other things a [saw] mill requires (…) even the processes by which the lacquer is made a beautiful yellow involve the skills of more persons than one can enumerate." (Read L., 1958, pp. 5, 6 & 7).

Accounting for the cumulative environmental impact of a given product system is an enormous task. It requires tracing and measuring the metabolism of all the activities associated with a particular product. Life Cycle Assessment (LCA) is one of the leading techniques that has been developed for this purpose (ISO, 2006).

1.1.2 LCA: Background data is key—but uncertain

LCA focuses on the "compilation and environmental evaluation of all processes, outputs and the potential environmental impacts associated with a product system" (ISO, 2006, p. 2). Its goal is to pinpoint ecological weaknesses, compare alternatives, evaluate the main environmental impacts, design new products, measure the environmental relevance of a material or product, and establish recommendations for actions (Guinée et al., 2002). In the context of environmental management, LCA is seen as a decision-support technique (Werner, 2002). Since 2000, LCA has received increasing attention from academics, policymakers, and the general public, and now plays a decisive role in business decisions related to sustainable production and in modern environmental policies (Düpmeier et al., 2012; European Commission, 2009; Guinée et al., 2011; Zah et al., 2009).

The LCA procedure is structured around a linear model, which views the product system as a network of basic building blocks called unit processes. A unit process represents one specific activity or group of activities in the product system, and records (i) the intermediate exchanges from and to the technosphere, that is, the input of usable energy and raw materials and the output of products and waste, and (ii) the exchanges with environment, that is, the input of natural resources and the output of emissions (Rebitzer et al., 2004). The unit processes are linked by means of their intermediate exchanges. The quantitative analysis of the product system results in the sum of all exchanges with the environment for the entire life cycle, called Life Cycle Inventory (LCI) (Rebitzer et al., 2004). The LCI is calculated with reference to a functional unit[1] that determines the product (or reference) flow of interest. Life Cycle Impact Assessment (LCIA)—the next step in the LCA procedure—then takes the inventory data on these exchanges as an input and uses it to quantify the potential impact on the environment.

Although the LCA procedure reduces the complexity of the real-world system[2], it remains a very data-intensive method (Finnveden et al., 2009; Hellweg and Milà i Canals, 2014). A typical product life cycle covers thousands[3] of unit processes, each of which needs to be described with exchange flow data. This information cannot be gathered within a specific project due to the high cost[4] of data collection. It is therefore common practice to focus data-collection efforts on selected activities that reflect the space for action—these activities are together called the "foreground system"—and use generic average data from process-based LCI databases to model the remaining activities—the "background system" (Bourgault et al., 2012; Tillman, 2000). In a typical LCA study, the background system covers up to 99% of the unit processes; only in rare cases does the number of unit processes modeled explicitly in the foreground system exceed 5%[5]. Background or LCI databases therefore form the backbone of any LCA study. They represent most of the building

[1] See Glossary in Chapter 6 for definitions of these and other key terms.

[2] It removes single-objective product systems from a multi-objective world. That is, the product system is a theoretical construct that has no clearly definable temporal or spatial boundaries in a complex context (Werner, 2002).

[3] In fact, the number of processes per product system expands in line with the size of background databases (Steubing et al., 2016).

[4] Collecting the approximately 10,000 datasets in the ecoinvent database, probably the largest of all LCI databases, took more than 15 years. This gives some indication of the tremendous effort required for managing LCI databases.

[5] Even if 100 processes are modeled with primary data, the foreground system will still not exceed 5% of the entire product system as it typically involves more than several thousand unit processes (Steubing et al., 2016).

blocks required for any LCA, that is to say, aggregated and/or disaggregated unit process datasets. The quantity and quality of unit process data-provided LCI databases is thus of the utmost importance.

However, the unit process data provided by LCI databases is typically uncertain and incomplete (Groen et al., 2014). For this, there are a number of reasons. First, it is not feasible to match each original activity in the real world with a description of a unit process in the LCI database, because the high cost of data collection typically allows only for generic, average datasets to be generated. Consequently, the exchange flow values recorded in a unit process often represent assumed average conditions for a whole country, or even a much larger region, across a given time period and across different instances of real processes. In reality, of course, the characteristics of industrial production and environmental conditions can vary significantly over geographic space, time, and process instances, and consequently natural variability is always present (Huijbregts, 1998). This fact limits the accuracy of LCA—particularly in the field of agriculture, where exchange flows are highly sensitive to the natural variability of the environment.

Second, the quantitative data required to compile an accurate representation of a specific activity may be unavailable, wrong, or unreliable (Ciroth et al., 2004; Heijungs and Huijbregts, 2004). Moreover, the growth of LCI databases during the last twenty years has primarily been funded by national institutions or specific research projects (Mutel, 2012). Consequently, LCI databases are biased toward the country where they originate in their coverage of real-world activities (Steubing et al., 2016). Although current developments improve on this shortcoming (see, for example, Wernet (2012)), today's LCI databases are still far from complete, even on the level of generic data. Due to this incompleteness, process-based LCI databases lack a significant proportion of the metabolism of our economic system required to fulfill any given product demand (Majeau-Bettez et al., 2011; Suh and Huppes, 2005, 2002). Standard LCA applications therefore systematically underestimate environmental impacts (Lenzen, 2000; Lenzen and Dey, 2000).[6]

[6] Different approaches have been developed to hybridize process-based LCI databases with Environmentally Extended Input-Output (EEIO) analysis (Suh and Huppes, 2005, 2002). EEIO analysis can be considered the top-down complement to process-based LCI databases. It is based on regional or national accounting tables that describe all activities in an economy at the level of sectors, and their corresponding use of resources and release of emissions (Majeau-Bettez et al., 2011). While such EEIO inventories cannot be used for detailed product-level LCA due to their lack of specificity, their completeness offers an interesting starting point for enhancing process-based LCAs (Finnveden et al., 2009; Majeau-Bettez et al., 2011; Suh and Huppes, 2002). However, hybrid analysis has "yet to enter mainstream practice" (Majeau-Bettez et al., 2011, p. 10171). Rather than focusing on enhancing LCI databases by using EEIO, this dissertation focuses on identifying key

To summarize, LCA is a very important but very data-intensive technique, which would be practically impossible without using the unit process datasets provided by LCI databases. Although the increased availability of background data from LCI databases has substantially decreased the effort of conducting an LCA, the high cost of data collection still imposes key limitations on the quality of LCA studies. These limitations either take the form of incomplete and uncertain background data, which must be gathered and maintained at great effort, or they result in a generic representation of highly context-dependent activities, for example, agricultural work—activities which "must focus on averages at the expense of specificity" (Mutel and Hellweg, 2009, p. 5802).

1.2 Area of investigation

This dissertation develops two complementary solutions based on information systems (IS) that increase the effectiveness and efficiency of data collection in LCA. The first solution is a statistical prioritization approach that directs LCI database improvement efforts toward the most important datasets in terms of their potential influence on overall database quality. This first approach aims to enhance the effectiveness of data collection using LCI databases. The second solution is a computational framework that makes it easier to automatically generate regionalized cultivation datasets on the basis of publicly available spatial (raster) data. This approach aims to increase the efficiency and quality of unit process dataset generation in the domain of agricultural LCAs. In the following sections (1.2.1 and 1.2.2), we explain why we believe that this area of investigation offers particular value to the field of LCA.

Throughout this dissertation, the ecoinvent LCI database is used to evaluate methods or compute case studies. With more than 10,000 datasets, the ecoinvent database is the world's largest LCI database. Its size and associated complexity (see Section 1.2.1), as well as the fact that it provides disaggregated unit process datasets, means that it is an appropriate basis for implementing (and then evaluating) the computational frameworks developed in this dissertation.

---

environmental issues in process-based LCI databases. The statistical prioritization approach developed in this dissertation could also be used to identify key data elements in EEIO or hybrid LCA tables.

1.2.1 Prioritizing LCI database improvement

We distinguish two perspectives for prioritizing improvements to the LCI database: external prioritization and internal prioritization. External prioritization aligns inventory effort[7] with importance based on an external reference, such as environmentally extended input-output (EEIO) databases. For example, Majeau-Bettez et al. (2011) contrast the economic, environmental, and structural importance of economic sectors (based on an EEIO database) with the number of corresponding unit processes in Version 2.1 of the ecoinvent database in order to identify the sectors that are most underrepresented. The discrepancy between the importance of different sectors in the EEIO database and the corresponding inventory effort leads Majeau-Bettez et al. to suspect a "suboptimal allocation of LCA inventory resources" (Majeau-Bettez et al., 2011, p. 10175).

Internal prioritization aligns data collection efforts by identifying key environmental issues in the existing data. To the best of our knowledge, LCI database improvements to date have not been systematically guided by identifying key elements, even though the identification of such elements in LCA studies is well developed (Heijungs, 1996; Heijungs et al., 2005; Heijungs and Kleijn, 2001).

Insights into relative process importance are becoming increasingly relevant due to the growing numbers of unit processes in existing LCI databases. For example, Version 3.1 of the ecoinvent database includes between 10,305 and 11,332 unit processes, depending on the system model[8]. If updating one unit process on average requires only one person day[9], a systematic update of the entire database within one year would require the continuous work of more than 50 researchers. Typically, capacity for such extensive improvement efforts is not available.

In addition, the unit process datasets in LCI databases are not autonomous. The mutual dependencies between datasets form a complex graph that can be considered a simple physical

[7] We use the term "inventory effort" to mean the process of generating, maintaining, or updating unit process datasets.

[8] A system model provides a set of rules that specify how activity datasets are linked to form product systems. Version 3.1 of ecoinvent offers three system models to choose from. They differ mainly with regard to their handling of the multi-functionality problem (system expansion vs. allocation), their use of average or marginal suppliers, and their assessment of by-product treatments (Weidema et al., 2013).

[9] The workload for updating a dataset varies greatly. Depending on the level of completeness of a dataset, it might take weeks (e.g., where a global dataset is disaggregated into many country-specific datasets) or just an hour (e.g., where the quantity of a single emission flow in a single dataset is updated). Updating a dataset typically involves data collection, data entry/manipulation, and data submission to peer review using the software tool EcoEditor. Any change to existing datasets or submission of new datasets requires peer review to be accepted into the database (Weidema et al., 2013). Therefore, one person day should be considered as a rough but realistic estimate of the average effort associated with updating a dataset.

network model of our modern economic system. Therefore, LCI databases represent highly interconnected systems, where almost every unit process is involved in every product system.

The average number of processes involved per product system is expanding in line with the size of the background database. For example, the average product system in Version 2.2 of the ecoinvent database (released in 2007) includes 2,426 unit process datasets, while Version 3.1 (released in 2014) already includes approximately 8,000 datasets (Steubing et al., 2016). As noted by Heijungs (2012, p. 172), "due to the interlinked nature of many processes, seemingly alien[10] parts of the database are involved in all LCAs". Without detailed insight into relative importance of different processes, the increasing size of LCI databases is making the effective organization of improvement efforts more and more difficult.

Moreover, improvements in the quality and quantity of data are usually the result of external data availability and situation-driven requirements rather than systematic choices (Sonnemann and Vigon, 2011). The continuous improvement of datasets and methodologies for LCA is driven by the "continuing evolution in consumer preferences, market and industry imperatives, and public policy" (Sonnemann and Vigon, 2011, p. 98). Compliance with these developments typically ties up much of the limited workforce, but is "probably not the most effective use of resources in improving overall database quality" (Mutel, 2012, p. 128). Overall, the impact of current data updates on the quality of LCI databases remains unclear, and it is possible that maintenance efforts are ineffective.

Given the cost of data collection, improving LCI databases should be organized as effectively as possible. We believe that insights into the relative importance of different processes in existing LCI databases are a crucial, currently unmet condition for organizing data-collection efforts more effectively. Knowing the relative importance of the different processes in an LCI database would make it easier to take the right action, improving those data elements that have the greatest influence on the overall quality of the database. We therefore formulate our first research question as follows:

[10] For example, the product system associated with the provision of one kilowatt hour (kWh) of high-voltage electricity produced by a hydroelectric power plant in the Swiss Alps involves third-party processes such as airport infrastructure, sugar (from sugar cane), kraft paper, and many more. This results from the high degree of interconnectedness between the unit process datasets in the database (Steubing et al., 2016).

RQ1: How can we identify and use relative process importance in LCI databases to organize data collection efforts more effectively?

1.2.2 Regionalized LCI modeling

In practice, LCA is mostly structured around the use of either site-generic[11] or site-dependent[12] unit process datasets; site-specific[13] or regionalized datasets are used very rarely (Mutel et al., 2012). It is common to consider the average meteorological and ecological conditions of a whole country and geographical region when compiling the intermediate exchanges and exchanges with the environment listed in the unit process dataset. The cost of data collection means that only rudimentary site-specific datasets can be generated. Usually, the LCA "must focus on average values at the expense of specificity" (Mutel and Hellweg, 2009).

This factor limits the representativeness and accuracy of LCAs—particularly in the field of agriculture, which involve exchange flows that are highly sensitive to the natural variability of the local environment. The type and amount of resources used (water, land, and so on), the intermediate flows required (for example, the application of mineral and organic fertilizer or the use of machinery), and the accompanying release of emissions into soil, air, and water (for example, nitrate, di-nitrogen monoxide, and phosphate) is determined by micro-spatial environmental parameters (precipitation, soil properties, slope, and so on) and is therefore are highly context-dependent. For such processes, even small changes in local bio-geographical conditions can alter the type and magnitude of the exchange flows included, and hence their environmental impact (Geyer et al., 2010a).

The "unit world" assumption almost inherent to LCA largely ignores this variability. Although explicit spatial data on micro-spatial conditions is now available, with good resolution and on a global scale (Hengl et al., 2014; Monfreda et al., 2008; Mueller et al., 2012), considering such data when generating unit process datasets is too labor-intensive[14] using normal methods of data

[11] "Site-generic" datasets represent an average for large geographic regions, such as continents or the entire globe (Mutel et al., 2012).

[12] "Site-dependent" datasets follow country or state boundaries (Mutel et al., 2012).

[13] In this dissertation, the term "site-specific" is used interchangeably with the term "regionalized" to mean that a site-specific dataset always refers to an individual location, such as a particular plot or plant (Mutel et al., 2012).

[14] Assuming that generating one dataset would require one day, generating a site-specific dataset for each grid cell in a resolution of 1 arc minute (approximately 10x10 km) for all grid cells in Germany would require approximately 5,000 days.

processing. Agricultural unit process datasets are mainly generated manually, according to specific guidelines[15] and emissions models and drawing on a wide array of raw data sources, ranging from public available databases (FAOSTAT, EUROSTAT, and so on) to company data, surveys, case studies, publications, measurements, and so on (Nemecek et al., 2015). In addition, processing spatial data into a comprehensive unit process dataset is by no means a straightforward task. Many of the spatial parameters require a high degree of manipulation and integration with each other until they represent one or several relevant exchange flows in an agricultural unit process dataset. On the other hand, because not all exchange flows can be computed on the basis of spatial data, datasets needs to be complemented with default data from background databases.

We believe that spatial data can increase the efficiency and quality of agricultural dataset generation in the domain of agricultural LCA when combined with a computerized method for regionalized LCI modeling. Regionalized LCI modeling is the procedure that generates and links process datasets to the location where they occur (Mutel et al., 2012). It is motivated by the "recognition that industrial production characteristics (…) vary throughout space" (Mutel et al., 2012).

The significance of regionalized LCI modeling for improving the representation of unit process datasets has been acknowledged in the past (Mutel, 2012; Mutel et al., 2012; Mutel and Hellweg, 2009; Seto et al., 2012). However, most of the literature on regional aspects in LCA focuses on regionalized LCIA modeling—recognizing that the location of a source and the conditions of its surroundings influence the environmental impact (Hauschild, 2006). To the best of our knowledge, regionalized LCI modeling in the field of agriculture has only been applied in six case studies (Dresen and Jandewerth, 2012; Geyer et al., 2010a, 2010b; Reinhard et al., 2011; Scherer and Pfister, 2015; Zah et al., 2012). All of these studies succeed in the case-specific parameterization of the spatial properties of selected exchanges. However, none of them provides a general framework for regionalized LCI modeling in LCA that provides comprehensive unit process datasets. We therefore formulate our second research question as follows:

RQ2: How can we automatically process publicly available spatial data into regionalized comprehensive unit process datasets that consider existing background data from LCI databases?

[15] Such as the data quality guidelines from ecoinvent (Weidema et al., 2013), the Agri-footprint database (Blonk, 2014), or the WFLDB (Nemecek et al., 2015).

1.2.3 Solution objectives

The overall goal of this dissertation is to develop computational frameworks for organizing the collection of data for LCAs more efficiently and effectively. We approach this goal as follows:

1. **Develop a calculation method for database-wide contribution analysis (CA).** Although some studies use CA for analyzing certain features of LCI databases, the use of CA has not been formalized with respect to prioritizing improvements to the LCI database. The calculation method we develop should compute the relative contribution of each unit process throughout each product system with respect to the full spectrum of selected LCIA[16] indicators. Relative contributions of this type should be calculated for two perspectives: the *causer* perspective and the *connector* perspective. These perspectives represent the two possible capacities of a unit process, causing and connecting to environmental impacts. They thus represent complementary tools that we can use to achieve our goal of identifying important processes throughout the database. In the causer perspective, we focus exclusively on the causing elements of each unit process, that is, the direct exchanges with the environment that cause environmental impacts. This perspective helps pinpoint those processes with consistently large contributions in terms of environmental interventions. In the connector perspective, we focus exclusively on the connecting elements of each unit process, that is, the intermediate flows from other processes within the technosphere. This perspective helps us pinpoint the unit processes that are consistently linked to large upstream contributions.

2. **Develop a set of summary measures to investigate the relative importance of different processes.** Summary measures "condense a certain aspect of a large set of numbers into one or perhaps a few items" (Heijungs and Suh, 2002, p. 152). Using the database-wide CA as a starting point, our set of summary measures (or "assessment metrics") should make it possible identify key datasets and summarize the inequality in the relative importance of different processes**.** Measuring this inequality of process importance—for example, how many processes account for how much of the overall contribution—provides important insights into the utility of the prioritization.

[16] In principle, CA can also be applied at the inventory level (Heijungs and Kleijn, 2001). However, we focus on environmental impacts as we believe that a CA at the inventory level is of little practical interest in the context of prioritizing improvement efforts.

3. **Develop a ranking algorithm to determine the importance of datasets across LCIA indicators**. The relative importance of a process is a function of the particular LCIA indicator applied, that is to say, different LCIA indicators will prioritize different processes. In other words, prioritizing for a single LCIA indicator would lead to a one-sided improvement of the LCI database. So for the prioritization to be robust, many LCIA indicators must be considered. The method used should allow us to calculate an overall ranking of unit processes considering their importance across any given set of LCIA indicators.

4. **Demonstrate and evaluate the prioritization method developed in a proof-of-concept application and a comprehensive analysis of an LCI database.** The proof-of-concept application should explore the utility of the method in its practical application, in terms of effectiveness. The comprehensive analysis of the ecoinvent database should test the method in its practical application and provide detailed insights into the structural dependencies in the database.

5. **Develop a regionalization method for automatically generating regionalized cultivation datasets on the basis of publicly-available spatial (raster) data**. The method should process publicly available spatial (raster) data into comprehensive unit process datasets using background data from the ecoinvent database and the emission models from the World Food Life Cycle Database (WFLDB) guidelines.

6. **Demonstrate and evaluate the regionalization method in a case study**. The case study should explore the utility of the method in its practical application.

This dissertation addresses these objectives in the form of three research articles. Articles I and II focus on objectives 1 to 4, while Article III focuses on objectives 5 to 6. We provide a brief summary of each article in Chapter 1.4.

1.3 Methodological approach

Our methodology is based on the Design Science Research (DSR) approach[17]. DSR is concerned with the systematic creation of knowledge by design. Its defining feature is "learning through building". At its core, it is about constructing an artifact that is intended to solve a problem. In

[17] For more detail on DSR, see, for example, (Gregor and Hevner, 2013; Hevner and Chatterjee, 2010).

accordance with Gregor and Hevner (2013, p. 341) we use the term "artifact"[18] to denote "a thing that has, or can be transformed into, a material existence as an artificially-made object (e.g., model, instantiation) or process (e.g., method, software)."

DSR is characterized by three design science research cycles (Figure 1-1). The *relevance cycle* ensures that the design cycle focuses on important opportunities and problems in an actual application domain. The *rigor cycle* connects the design cycle activities with the knowledge base of scientific foundation, that is, the state-of-the-art experience and expertise in the relevant application domain (Hevner and Chatterjee, 2010). Repeated grounding in the knowledge base ensures that the research represents an actual improvement. The *design cycle* switches back and forth between the core activities of building and evaluating the design artifact (Hevner and Chatterjee, 2010).

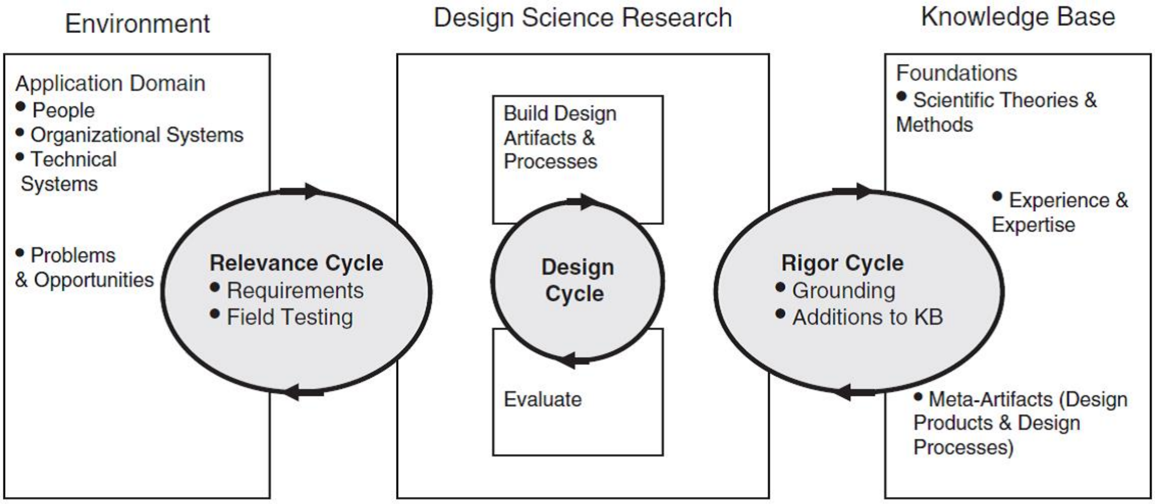


Figure 1-1: Design science research cycles. KB = knowledge base. Source: (Hevner and Chatterjee, 2010, p. 16)

Our design procedure follows Peffers et al. (2007) and is a step-by-step procedure for building artifacts and has found wide adoption in practice (Gregor and Hevner, 2013). It includes the following steps: (1) identify the problem, (2) define the objectives, (3) design and develop (build), (4) demonstrate, (5) evaluate, and (6) communicate (Peffers et al., 2007). Figure 1-2 shows the core activities associated with these steps and how we addressed them for both artifacts. In Section 1.2 we discussed steps (1) and (2), and then explained the key characteristics of the design cycle, steps (3) to (5). This dissertation and the associated publications represent step (6) in the procedure: communicate.

[18] We use the term "computational framework" interchangeably with "artifact".

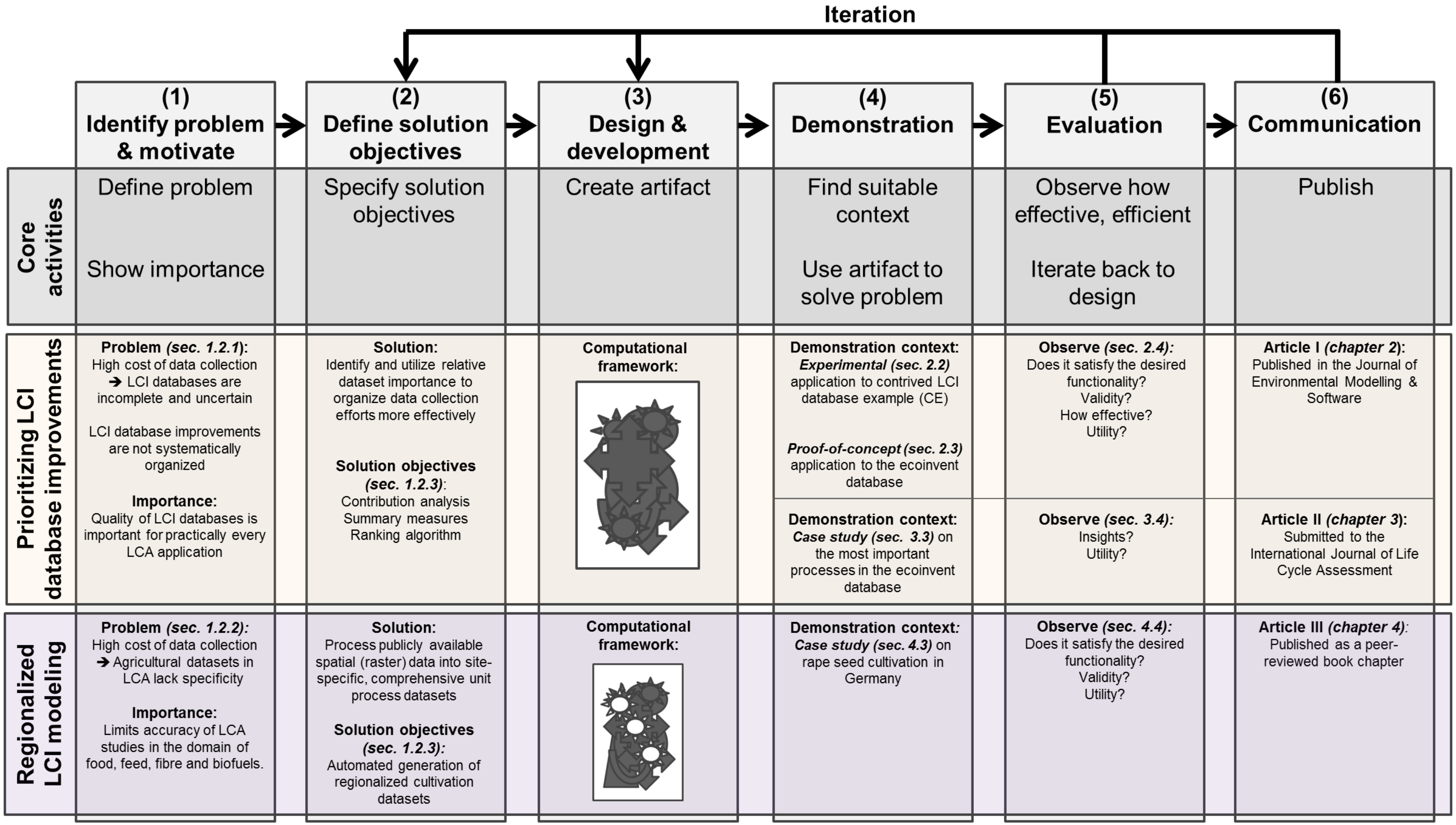


Figure 1-2: Design procedure and associated core activities for developing two artifacts: a computational framework for prioritizing LCI database improvements, and a computational framework for regionalized LCI modeling. For each step, we highlight how the core activities were addressed for our artifacts. "Sec." refers to the section of this dissertation where we provide more information. The fourth and the sixth solution objectives are not shown in step (2) because they form an integral part of the design procedure and are thus performed in steps (4) and (5). Adapted from Peffers et al. (2007, p. 54).

As mentioned above, the core feature of DSR is "learning through building". Consequently, the actual design procedure for each computational framework involves several iterations, particularly between the develop (build), demonstrate, and evaluate steps—steps (3), (4), and (5) in Figure 1-2.

We developed most features of the prioritization framework in an experimental setting (Article I), that is, we determined a particular feature of the artifact (3) first in its application to a "contrived example" (CE). The CE reflects important characteristics of the original database in a greatly simplified form (see Section 2.2.1). The application of the artifact to the CE demonstrated—step (4)—its utility with regard to the specific feature, and allowed us to immediately evaluate it—step (5). We repeated this first design cycle until the artifact sufficiently satisfied the intended feature for the CE. We then applied the artifact to Version 3.1 of the ecoinvent database (proof-of-concept)—step (4)—to observe how well it worked under real-world conditions—step (5). The artifact was then further improved until it fully satisfied the specific functional requirements for the feature of interest. We repeated the design cycle until the developed artifact sufficiently addressed the first and second solution objectives. We added a ranking algorithm (third solution objective) to the prioritization framework, and demonstrated and then evaluated the entire framework in its business environment with a comprehensive case study focused on identifying the most important processes in the ecoinvent database (Article II).

We developed the regionalization framework by continuously demonstrating and evaluating the anticipated features in a case study on rape seed cultivation in Germany (Article III).

1.4 The articles presented in this dissertation

The main content of this thesis is a collection of three research articles. Articles I and III were published in international peer-reviewed journals. Article II is based on a peer-reviewed conference paper that received the best paper award, and has been submitted to a book in the area of environmental informatics and LCA. All three articles deal with the development of computational frameworks that increase the effectiveness and efficiency of data collection in LCA. Below, we provide a brief summary of the articles and their focus.

1.4.1 Article I: Contribution-based prioritization of LCI database improvements. Method design, demonstration and evaluation

Article I examines the various facets of a computational framework that facilitates the systematic identification of key datasets in existing LCI databases, including a calculation method for database-wide contribution analysis (CA), summary measures to investigate key datasets and the inequality in their importance, the interpretation of prioritization results obtained from a proof-of-concept application to Version 3.1 of the ecoinvent database, and an evaluation of the utility of the proposed framework.

The article presents a contribution-based approach that helps organizations maintaining LCI databases identify which data elements should be prioritized for improvement. Using a contrived LCI database example, we present a novel workflow for computing database-wide contribution analysis in a formalized yet general manner. For a particular LCIA indicator, the workflow yields two relative contribution matrixes: one that indicates the direct (causing) and one that indicates the upstream (connecting) relative contribution of each unit process throughout each product system in the database. These perspectives represent complementary lenses for our goal of identifying important processes throughout the database.

We then turn to the question of how the information in these matrixes can be investigated. Besides the arithmetic mean, which is required to summarize relative process importance across the relative contribution matrixes, we examine the Lorenz curve—the most popular tool for comparing income and wealth inequality in the field of economics (Duclos and Araar, 2006)—and associated measures of inequality, such as the Concentration ratio and the Gini coefficient, for assessing concentration and inequality in relative process importance. The points on the Lorenz curve represent statements such as "the bottom 50% of the households have only 20% of the total income", and provides some interesting properties for analyzing and assessing the inequality in relative process importance.

We also demonstrate the method in its (proof of concept) application to Version 3.1 of the ecoinvent database for two selected LCIA indicators. The analysis reveals a remarkable disparity in relative process importance; 11 (out of 11,000 potentially available) processes cause more than 60% of the total relative contribution associated with photochemical oxidation (one LCIA indicator) throughout the database.

Finally, we evaluate our method based on the insights gained in the proof-of-concept application with regard to validity (does the method do what it is meant to do?), performance (how much time is required to perform the computations?), and utility (does the method provide value?). We conclude with a discussion of further potential areas of application and the limitations of the approach.

1.4.2 Article II: Contribution-based prioritization of LCI database improvements. The most important unit processes in ecoinvent

Article II examines the various facets of a computational framework that facilitates the systematic identification of key process datasets in existing LCI databases in its practical application to Version 3.1 of the ecoinvent database. It includes a brief review of calculation method for database-wide contribution analysis (CA), a ranking algorithm to determine dataset importance across many LCIA indicators, and an interpretation of the prioritization results obtained from a comprehensive analysis of the ecoinvent database.

We apply the prioritization method demonstrated in Article I to Version 3.1 of the ecoinvent database. We first focus on the assessment of three selected LCIA indicators in order to establish a basic understanding and highlight key characteristics of the approach. The inequalities in relative process importance are remarkable across all LCIA indicators: for example, just three datasets cause 11% of all climate change impacts. We then present a novel consolidation algorithm based on the Lorenz curve that allows for the classification of relative unit process importance across any set of LCIA indicators.

We then use the consolidation algorithm to identify the most important unit processes according to a set of 19 selected LCIA indicators. The analysis shows that, even across many LCIA indicators, a relatively large degree of the overall database quality is dependent on a small set of key processes. Overall, 300 (out of roughly 11,000) datasets cause 60% of the environmental impacts across all 19 LCIA indicators and all product systems in the database.

We use the final ranking of processes to identify the sectors and geographies of particular importance. We show that processes related to electricity generation, waste treatment activities, and energy carrier provision (petroleum and hard coal) consistently have a major environmental impact in all product systems and across all the assessed LCIA indicators. In addition, our analysis highlights the presence and importance of central hubs, that is, sensitive intersections in the database network, whose modification can affect a large degree of the database quality. Electricity generation in particular, but also iron and steel production processes, have strong networking effects. We discuss several reasons for the major disparity in process importance and the potential implications for database management. We conclude with a discussion of the limitations of the approach and related future work.

1.4.3 Article III: Regionalized LCI modeling. A framework for the integration of spatial data in Life Cycle Assessment

Article III explores the various aspects of a computational framework that uses publicly available spatial (raster) data to automatically generate regionalized cultivation datasets. The article includes design principles for generating regionalized data structures, a case study on rape seed cultivation in Germany, and an initial examination of the significance of further use cases.

The article presents a computational framework for regionalized LCI modeling that allows the automated generation and assessment of regionalized unit process datasets. We transform publicly available spatial (raster) data into comprehensive unit process datasets using background data from Version 3.2 of the ecoinvent database and the emission models from the World Food Life Cycle Database (WFLDB) guidelines (Nemecek et al., 2015). The framework operates on the basis of a compiled repository of spatial raster data indicating harvested area, yield, fertilizer application rates, and so on, for all major crops, as well as data on precipitation, soil properties, and terrain. Using the latitude-longitude combination as a unique index, we merge raster datasets of different resolution and formats into a single parameter table. Each column of the parameter table represents, for a given latitude-longitude combination, the corresponding grid cell values of all raster datasets. We match, transform, and expand selected parameter table entries to exchange flows from the ecoinvent database in order to generate a comprehensive unit process dataset. The inventory table already represents comprehensive regionalized cradle-to-gate agricultural unit process datasets. We then generate a vector of impact factors for each exchange flow in the inventory table and assess the environmental impacts of the regionalized inventory.

Next, we demonstrate the framework in a case study on rapeseed cultivation in Germany. Using a resolution of 30 arc seconds (approximately 1x1 km), we generate and assess around 580,000 regionalized cradle-to-gate unit process datasets for rapeseed cultivation. The assessment reveals substantial spatial variation in environmental impacts. Contribution to climate change varies by a factor of two, while marine and freshwater eutrophication varies by a factor of five. Much of this variation is related to the variability in nitrate, phosphorous, and N_2O emissions. This confirms that the spatially-explicit computation of these flows is important in order to obtain accurate cultivation process datasets. Finally, we compare the size of key exchange flows with data from the literature and discuss the limitations of the study and five promising use cases for the framework presented.

1.5 Main achievements and impacts

Given the expense of primary data collection, precise knowledge of relative process importance is a prerequisite for effectively improving LCI databases. We have developed a statistical prioritization framework that helps organizations maintaining LCI databases identify the key processes to be prioritized for improvement. Demonstrating and evaluating the prioritization framework using the ecoinvent database shows that this is a useful approach: We provide evidence that a robust, tiny set of unit processes is responsible for most of the contribution across the entire LCI database, and even across the broad spectrum of all modern LCIA indicators (Article I, Section 2.4.5). Concentrating research efforts on increasing information density in these processes of systemic importance offers a new starting point for systematically and effectively improving the entire database.

Our ranking of the most important processes strengthens the basis for decision-making used by those managing the data, who can now systematically process the list depending on the time and resources available and decide for each unit process if and how it should be improved (Article II). Our prioritization framework thus forms a basis for effectively directing data collection efforts. Indeed, one noteworthy impact of our research is that ecoinvent database management team have decided to apply our method to developing priorities for updating and maintaining the ecoinvent database.

By implementing a computational framework for regionalized LCI modeling (Article III), we show that automatically generating regionalized cultivation datasets harbors great potential for improving the quantity and quality of agricultural unit process datasets in LCA. With 580,000 datasets, the case study presented is likely the most comprehensive cradle-to-gate LCA on climate change and the eutrophication impact of rapeseed cultivation in Germany. Highly important and spatially-sensitive exchange flows such as N_2O, nitrate, and phosphorous are explicitly computed for each grid cell (approximately 1x1 km) using a wide array of spatial-specific input parameters, such as precipitation, soil organic carbon content, erodibility, length-slope factor, and so on. This makes it easier to consider spatial conditions that are not generally taken into account when generating agricultural datasets, where the focus is typically on country averages.

Our regionalization framework is applicable to all regions of the world and all major crops. It thus forms a basis for efficient, geographically representative generation of high-resolution agricultural process datasets. The framework will be used to generate bottom-up computed average datasets for cotton cultivation—an agricultural commodity for which there are currently no representative cultivation datasets—for the World Apparel and Footwear Life Cycle Assessment Database (WALDB).

1.6 Main contributions

1.6.1 Computational framework for prioritizing LCI database improvements

By developing, demonstrating, applying, and evaluating a computational framework for prioritizing LCI database improvements, this dissertation contributes a number of novel insights in the domains of LCA and information systems.

By developing, demonstrating, and evaluating a new IS-based solution (the prioritization framework) to an established problem (the high cost of data collection), this dissertation also contributes new design knowledge to the domain of information systems. This design knowledge includes the description of a mathematical *method* for database-wide contribution analysis for two complementary perspectives (see Sections 2.1.4 and 2.2.2), new *constructs* in the form of summary measures (such as the h index, which indicates how many datasets account for how much of the total contribution throughout the database for a particular LCIA indicator, or the distinction between the causing and connecting characteristic of a unit process; Section 2.2.3), a ranking *method* for prioritization across many LCIA indicators (Section 3.2.2), *design principles* in the form of a description of the sequence of steps between these elements (Sections 2.3.1 and 3.2.1.2), and an *instantiation* in the form of an operational software tool.

According to the criteria presented by Gregor and Hevner (2013, p. 340), these contributions can be considered a "Nascent design theory". The description of our prioritization framework at this level of abstraction—as operational principles—allows it to be operationalized "in a number of other unstudied contexts, thus greatly increasing the external validity of the research" (Gregor and Hevner, 2013, p. 341). For example, with this level of abstraction, our framework could be transferred to any system-modeling problem where interacting elements of the system can be stated as a set of linear equations and the question arises of which of these elements should be improved with the highest priority. This type of problem can occur in material flow modeling (Bornhöft et al., 2016), EEIO analysis (Majeau-Bettez et al., 2011), or integrated hybrid analysis (Suh and Huppes, 2005), for instance.

Regarding the contribution of this dissertation to the field of LCA, our research focuses attention on the significance of LCI database analysis. We show that detailed insights into relative process importance in existing LCI databases provide a valid foundation for more effectively directing data-collection efforts. The LCA community could benefit greatly from this insight. The administration team of every LCI database[19] could organize their data collection strategy more effectively using our prioritization framework, and so improve the quality of the key datasets that build the very foundation of practically every LCA application.

The methods developed in this dissertation can inform LCA research in many ways. Our method of database-wide contribution analysis could be used to improve performance in dealing with all the issues concerned with the relative environmental relevance of particular processes, economic sectors, or product categories in LCI databases. For example, Frischknecht et al. (2007) analyzes the relative environmental relevance of capital goods throughout the ecoinvent database to decide whether or not they can be excluded in standard LCA applications. Such questions could be addressed very quickly on the basis of the relative contribution matrixes, or alternatively the relative contribution matrixes could provide a consistent basis for topological network analysis.

Although generating such weighted networks is considered straightforward for dependency analysis of the interacting parts of a system, other researchers have "not yet been able to identify a logical and consistent way to do so" (Heijungs, 2015, p. 162). The relative contribution in our matrixes could represent the missing "common currency" required for this task. Our ranking algorithm identifies and prioritizes relative process importance across many LCIA indicators. The algorithm could be used to identify key process datasets in every standard LCA application.

Furthermore, thanks to its comprehensive analysis of the ecoinvent database, the research presented here deepens our understanding of the structural dependencies in one of the world's largest LCI databases. We provide novel summary measures for assessing the inventory support of available LCIA indicators and important differences between them (see Table S1-1 in Section 3.7.3.1 and the associated discussion). This highlights a new starting point for systematic plausibility checks and error analyses. We also offer an initial analysis that elaborates and assesses possible reasons for the remarkable inequality in relative process importance (see the discussion in Section 3.4). This draws attention to some key blank spots in the ecoinvent database. We contribute practical measures to utilize the high concentration of relative database importance to, for example, improve dataset quality where it matters the most, namely datasets of systemic importance.

[19] This applies to any LCI database that can be transformed into or is available in the form of a square technosphere matrix and a corresponding biosphere matrix. A characterization matrix is also required.

1.6.2 Computational framework for regionalized LCI modeling

We present a computational framework that builds a bridge between spatial data and LCA. The chief results of the research are: (i) a framework for generating and assessing comprehensive regionalized data structures, and (ii) a case study that shows the potential relevance of spatial data in the field of LCI modeling. This contributes novel insights to the domains of information systems and LCA.

By developing and demonstrating a new IS-based solution (the regionalization framework) to an established problem (the fact that the high cost of data collection limits the accuracy of unit process datasets in the field of agricultural LCA), our dissertation contributes new design knowledge to the field of information systems. This design knowledge includes the description of a *design principle* (Section 4.2.1)—in other words, a technical recipe—for processing publicly available spatial (raster) data into comprehensive unit process datasets using background data from Version 3.2 of the ecoinvent database and the emission models from the WFLDB guidelines, plus an *instantiation* in the form of case study (Section 4.2.2). Although the design knowledge is less formalized than our prioritization approach, it contributes new insights into how key technical features of the framework are realized, such as the generation of comprehensive agricultural unit process datasets on the basis of background data[20] or the capacity to process different raster data resolutions and formats. These design principles can be used to build other bridges in the field of LCA. For example, with the appropriate spatial explicit data on metallic and non-metallic mineral resource extraction from the U.S. Geological Survey (USGS, 2016) and suitable inventory models, our new approach could be used to generate regionalized process datasets on oil sands and various forms of metal mining. On a more abstract level, our design principles could be used to process and expand company data, exported from an enterprise resource planning system (ERP) into comprehensive unit process datasets. In this way our design principles could improve the continuous assessment and monitoring of organizational environmental performance in the context of an Environmental Management System (EMS).

Our research contributes to the domain of LCA by focusing attention on the potential relevance of spatial data in the field of LCI modeling. The case study shows that regionalized LCI modeling matters. Emissions of great environmental relevance, such as N_2O, nitrate, and phosphorous, also

[20] Even the most recent and advanced studies in the field of the spatially-explicit GHG assessments of crop production appear to neglect the background system and therefore an important part of the environmental impact (Carlson et al., 2016; Gerber et al., 2016).

show great spatial variability and can be significantly higher than the emissions recorded in generic inventories (see Figure 4-2 and Table 4-3 in Section 4.3). Our framework makes it easier to efficiently generate datasets for all major crops and all regions of the world, while simultaneously increasing at least two data quality[21] characteristics of agricultural process data. The framework increases the accuracy—or, more specifically, the geographical representativeness—of agricultural datasets by making it possible to consider spatial conditions that cannot be accounted for in site-dependent datasets. Furthermore, it can increase reproducibility by enforcing the consistent use of assumptions and methods, a topic of particular importance with regard to emission modeling[22]. Finally, the framework can increase completeness because it allows all the relevant data points for a particular spatial scale of interest to be calculated and considered.

Our framework also provides new possibilities for aggregating and analyzing agricultural process data. In this regard, it could be used as a tool to improve agricultural dataset coverage and representation in LCI databases by adding more geographically representative, bottom-up computed average datasets and corresponding spatial variability of exchange flows[23]. The aggregation could be performed on the basis of political boundaries, such as a particular country, state, or city district, or on the basis of relatively homogeneous regions. Moreover, the framework could be used as a test model for analyzing and evaluating advanced emissions models developed outside the LCA context. This, in turn, could improve the accuracy of the emissions recorded in agricultural unit process datasets and support consensus-building on the utility of different emissions models in agricultural LCA.

By explicitly linking datasets to the location where they occur, our research also increases the utility of LCA for other important environmental system analysis tools, such as the Environmental Impact Assessment (EIA)[24] (Finnveden and Moberg, 2005). Our framework may be relevant in all project settings where assessing the environmental impact of spatially-sensitive processes is key. For example, it could be relevant where a large-scale bioenergy producer wishes to compare the environmental performance of bioenergy plantations in an explicit spatial setting. In this context, it

[21] See Glossary in Chapter 6 for a definition of important data quality aspects.

[22] Due to the inconsistent use of assumptions and methods, recent updates to the ecoinvent database have focused on harmonizing the emissions modeling in agricultural datasets (Nemecek et al., 2014).

[23] At present, exchange flow specific variability in LCI databases (used for Monte-Carlo analysis) is largely based on rough estimates.

[24] EIA is a procedural tool used for assessing the environmental impact of a project in an explicit spatial setting (Finnveden and Moberg, 2005).

offers an improved basis for decision-making as it makes it possible to explicitly consider variations in micro-spatial conditions that are otherwise ignored due to the common "unit world" paradigm of LCA. Alternatively, with the appropriate spatial data, it could improve the EIA of the Delhi-Mumbai corridor (DMIC, 2017), for example—a 90-billion-dollar infrastructure project.

1.7 Limitations and future work

1.7.1 A computational framework for prioritizing LCI database improvements

No direct support for the identification of blank spots: The relative process relevance is determined by the interconnection of available product systems. Our contribution-based approach is useful for elucidating the inherent structural dependencies in the existing network structure. Although this approach offers some indirect support for identifying blank spots—specifically, by identifying processes that are too generic and which should be modeled more specifically in terms of their spatial or technological scale—this inward-oriented perspective provides no direct support for identifying blank spots and is therefore unable to direct research efforts toward economic sectors that may be underrepresented. Future research should therefore complement our inward-oriented perspective with more outward-oriented prioritization methods, such as that of Majeau-Bettez et al. (2011). In this regard, the approach developed by Majeau-Bettez et al. (2011) would provide guidance for accumulating new unit process data, whereas our approach focuses on which elements within the present data structure should receive the most attention.

Simple summary indicators to reflect complex issues: Equating inventory effort with the number of unit processes assumes uniform data collection challenges for all processes (see Figure 2-13 and Figure 3-4). Although this ignores the highly heterogeneous nature of unit processes, we expect, following Majeau-Bettez et al. (2011, p. 10175), that "the great number of processes in the ecoinvent database allowed for an averaging out of extremes and a reliable 'average inventory effort' indicator". It should also be noted that our measure of benefit—the cumulated contribution reviewed—is not guaranteed to quantify the actual improvement potential (see Figure 2-13). Identifying the actual improvement potential would require assessing dataset quality, either by reviewing key datasets manually or by considering data quality metrics. Future research should focus on developing and integrating dataset quality metrics, such as spatial sensitivity and dataset uncertainty (Heijungs and Kleijn, 2001). This would support a much more fine-grained configuration of improvements and so further improve the effectiveness of the computational framework.

Discretization to size classes comes at the cost of information loss: Our consolidation algorithm enables the ranking of unit process importance across any set of LCIA indicators. The sorting procedure maintains essential information about the actual size of a unit process contribution because the rank of a process is first determined by its actual size class—the threshold that the process helps to exceed (see Section 3.2.2)—and only then by its LCIA support, which indicates how many LCIA indicators point to the process in question. This ensures that the processes required to exceed the first threshold, only prioritized by one, rather uncorrelated LCIA indicator with a low LCIA support (such as agricultural land occupation or ionized radiation), will still receive more attention—that is, a higher ranking—than a process with high LCIA support in the second threshold (see Table 3-3 in Section 3.3.2.1). However, the discretization to size classes comes at the cost of information loss. In particular, large contributions are reduced to the size class of the threshold. This reduces the possibilities of assessing overall inequality in a single number with the Gini coefficient. Future research should examine the performance of our approach and compare it with alternative ranking procedures.

Analyze the inventory support of different LCIA indicators: Our summary measures provide new starting points for assessing the inventory support of available LCIA indicators and key differences between them (see Table S1-2 in Section 3.7.3.1). We suggest that future research activities should analyze different LCIA indicators paradigms, for instance, a set of midpoint methods versus a set of endpoint methods, and the corresponding differences in their recommendations of key datasets. This would reveal the detailed inventory support of different LCIA paradigms and so offer important feedback for developers of the LCIA method and the LCI database.

1.7.2 Computational framework for regionalized LCI modeling

LCA is not regionalized: Our computational framework does not represent a fully regionalized LCA. Such an LCA would operate on the basis of a fully regionalized product system, with each process linked to the location where it occurs, and a regionalized or spatially differentiated LCIA method that recognizes that the location of emissions and the conditions of its surroundings influence the environmental impact (Hauschild, 2006; Heijungs, 2012; Mutel, 2012). Our framework does not operate on the basis of fully regionalized product systems. In fact, it only regionalizes the exchange flows in the process dataset—our foreground system—and takes the impact factors associated with these exchange flows, without further adaptation, from the ecoinvent database[25]. In other words, we do not assess the environmental impact of our geo-referenced emissions in the explicit spatial setting. Regionalized LCIA could be implemented either by integrating and matching regionalized characterization factors, such as that proposed by Helmes et al. (2012) for freshwater eutrophication, to our geo-referenced emissions, or by computing characterization factors directly in the framework. The latter solution would require integrating fairly advanced fate models, but would allow a spatially explicit characterization factor to be generated in the resolution of the inventory datasets. Future research should investigate the complexity and utility of both possibilities in more detail.

Quality of the spatial data: The framework is built upon spatial raster data, a fairly new source of raw data in the domain of LCI modeling. The use of spatial raster data generates dependencies, but it also creates new opportunities. For example, spatially explicit data on crop production and fertilizer input is, to our knowledge, only available from EarthStat (Mueller et al., 2012). This means that the results of our regionalization framework are currently bound to the year 2000, and therefore subject to future-dependent updates. Recent initiatives for open spatial data (Earth Observation Center, 2017; FAO, 2017; Panagos et al., 2015, 2014) may reduce such dependencies in the long term. On the other hand, many of the spatial raster files in our repository come with a spatial-explicit rating for data quality (Mueller et al., 2012). This rating is not used in the current framework. Future work should therefore focus on integrating such data quality ratings so that uncertainty can be assessed.

No one-size-fits-all-resolution: We computed our case study in a grid cell resolution of 30 arc seconds (approximately 1x1 km). This may seem like an excessive level of detail, but in fact it is still not sufficient to correctly model some of the exchange flows. For example, phosphorous emissions are mainly affected by the length-slope (LS) factor, and nitrate emissions by the erodibility factor. These spatial factors can vary greatly, even in a resolution of approximately 20x20 m. While spatial data for these factors would be available in such resolutions (Panagos et al., 2015, 2014), it would be computationally inefficient to compute all the exchange flows in this resolution, especially as most of the spatial input data follows a grid-cell resolution of approximately 10x10 km (see Table 4-1 in Section 4.2.1). Future research should therefore examine a tiered calculation approach that

[25] We use the default (archetypal) compartment (air, water, soil) and sub-compartment (high population, low population, indoor, etc.; river, ocean, etc.) classification system provided by the ecoinvent database to determine the appropriate characterization factor of an emission.

adapts the resolution of the calculation to the relevance and spatial sensitivity of the exchange flow. The first step toward creating such an approach would be a detailed investigation of the spatial sensitivity of all important exchange flows.

Aggregation of regionalized unit process datasets: Many applications of regionalized datasets will require those datasets to be aggregated into representative averages. For example, the framework can be used as a tool to improve LCI databases by adding more accurate average agricultural unit process datasets. Consistent, spatially-explicit, bottom-up computation of unit process datasets would make it possible to generate more accurate averages for relatively homogeneous regions. However, what constitutes a homogeneous region "is a matter for scientific inquiry" (Mutel, 2012, p. 1), meaning that the ideal spatial scale of a unit process dataset depends on focused exchange flow. For example, the spatial scale of a unit process dataset optimized for the reduction in variability of nitrate emissions will be much smaller than the spatial scale of a unit process datasets optimized for the reduction of variability in fertilizer application rate (see the coefficient of variation in Table 4-3 in Section 4.3). Future research should therefore investigate multi-objective aggregation procedures that feature trade-offs between exchange flow relevance (in terms of environmental impact), variability, data quality, and spatial proximity to build representative unit processes datasets from the body of data produced by our framework. This would increase the representativeness of agricultural datasets in LCI databases and improve the general utility of the framework for the domain of LCA. The performance of these approaches should be compared with the spatial autocorrelation approach proposed by Mutel et al. (2012).

Incomplete emission models: The emission models (Nemecek et al., 2015, p. 35 Table 5) used in agricultural unit process datasets are often adjusted to the "averaging" nature of LCA, that is, they lack important compartments due to their generic spatial and temporal orientation. For example, the computation of nitrate ignores losses with waterborne and windborne sediment. However, as our results section shows (Section 4.3), the emissions in agricultural unit process datasets often contribute significantly to the total environmental impact. Our framework offers a modular and expandable test model for the application of advanced models. Future research should focus on evaluating more advanced emission models developed in areas other than LCA.

1.8 References

Blonk, C., 2014. Agri-Footprint - Part 2 - Description of data - Version 1.0.

Bornhöft, N.A., Sun, T.Y., Hilty, L.M., Nowack, B., 2016. A dynamic probabilistic material flow modeling method. Environ. Model. Softw. 76, 69–80. doi:10.1016/j.envsoft.2015.11.012

Bourgault, G., Lesage, P., Samson, R., 2012. Systematic disaggregation: a hybrid LCI computation al-gorithm enhancing interpretation phase in LCA. Int. J. Life Cycle Assess. 17, 774–786. doi:10.1007/s11367-012-0418-7

Carlson, K.M., Gerber, J.S., Mueller, N.D., Herrero, M., MacDonald, G.K., Brauman, K.A., Havlik, P., O[rsquor]Connell, C.S., Johnson, J.A., Saatchi, S., West, P.C., 2016. Greenhouse gas emissions intensity of global croplands. Nat. Clim Change advance online publication.

Ciroth, A., Fleischer, G., Steinbach, J., 2004. Uncertainty calculation in life cycle assessments. Int. J. Life Cycle Assess. 9, 216–226. doi:10.1007/BF02978597

DMIC, 2017. ::: Welcome to Delhi Mumbai Industrial Corridor ::: [WWW Document]. URL http://delhimumbaiindustrialcorridor.com/ (accessed 1.18.17).

Dresen, B., Jandewerth, M., 2012. Integration of spatial analyses into LCA—calculating GHG emissi-ons with geoinformation systems. Int. J. Life Cycle Assess. 17, 1094–1103. doi:10.1007/s11367-011-0378-3

Duclos, J.-Y., Araar, A., 2006. Poverty and Equity: Measurement, Policy, and Estimation with DAD. Springer.

Düpmeier, C., Greceanu, C., Kusche, O., Schmitt, C., 2012. Managing LCI Data from Different Work-groups within the same Instance of an LCA Database., in: EnviroInfo Dessau 2012, Dessau, Germany, August 29-31, 2012. pp. 507–515.

Earth Observation Center, 2017. . Copernic. Eur. Erdbeobachtungsprogramm. URL http://www.dlr.de/eoc/desktopdefault.aspx/tabid-5367/9013_read-16792/

EC-JRC, 2011. International Reference Life Cycle Data System (ILCD) Handbook-Recommendations for Life Cycle Impact Assessment in the European context. First edition November 2011. EUR 24571 EN. Luxemburg. Publications Office of the European Union;

European Commission, 2009. Directive 2009/125/EC of the European Parliament and of the Council of 21 October 2009 establishing a framework for the setting of ecodesign requirements for energy-related products.

FAO, 2017. Agro-Maps. Global Spatial Database of Agricultural Land-use Statistics [WWW Document]. URL http://kids.fao.org/agromaps/

Finnveden, G., Hauschild, M.Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Penning-ton, D., Suh, S., 2009. Recent developments in Life Cycle Assessment. J. Environ. Manage. 91, 1–21. doi:10.1016/j.jenvman.2009.06.018

Finnveden, G., Moberg, Å., 2005. Environmental systems analysis tools – an overview. J. Clean. Prod. 13, 1165–1173. doi:10.1016/j.jclepro.2004.06.004

Frischknecht, R., Althaus, H.-J., Bauer, C., Doka, G., Heck, T., Jungbluth, N., Kellenberger, D., Neme-cek, T., 2007. The environmental relevance of capital goods in life cycle assessments of pro-ducts and services. Int. J. Life Cycle Assess. 12, 7–17.

Gerber, J.S., Carlson, K.M., Makowski, D., Mueller, N.D., Garcia de Cortazar-Atauri, I., Havlík, P., Herrero, M., Launay, M., O'Connell, C.S., Smith, P., West, P.C., 2016. Spatially explicit esti-mates of N2O emissions from croplands suggest climate mitigation opportunities from im-proved fertilizer management. Glob. Change Biol. 22, 3383–3394. doi:10.1111/gcb.13341

Geyer, R., Lindner, J.P., Stoms, D.M., Davis, F.W., Wittstock, B., 2010a. Coupling GIS and LCA for biodiversity assessments of land use. Int. J. Life Cycle Assess. 15, 692–703. doi:10.1007/s11367-010-0199-9

Geyer, R., Stoms, D., Lindner, J., Davis, F., Wittstock, B., 2010b. Coupling GIS and LCA for biodiversi-ty assessments of land use. Int. J. Life Cycle Assess. 15, 454–467. doi:10.1007/s11367-010-0170-9

Gregor, S., Hevner, A.R., 2013. Positioning and Presenting Design Science Research for Maximum Impact. MIS Q. 37, 337–355.

Groen, E.A., Heijungs, R., Bokkers, E.A.M., de Boer, I.J.M., 2014. Methods for uncertainty propagation in life cycle assessment. Environ. Model. Softw. 62, 316–325. doi:10.1016/j.envsoft.2014.10.006

Guinée, J.B., Gorrée, M., Heijungs, R., Huppes, G., Kleijn, R., de Koning, A., van Oers, L., Sleeswijk, A.W., Suh, S., de Haes, H.A.U., de Bruijn, H., van Duin, R., Huijbregts, M.A.J., 2002. Hand-book on Life Cycle Assessment. Operational Guide to the ISO Standards.

Guinée, J.B., Heijungs, R., Huppes, G., Zamagni, A., Masoni, P., Buonamici, R., Ekvall, T., Rydberg, T., 2011. Life Cycle Assessment: Past, Present, and Future. Environ. Sci. Technol. 45, 90–96. doi:10.1021/es101316v

Hauschild, M., 2006. Spatial Differentiation in Life Cycle Impact Assessment: A decade of method development to increase the environmental realism of LCIA. Int. J. Life Cycle Assess. 11, 11–13. doi:10.1065/lca2006.04.005

Heijungs, R., 2015. Topological Network Theory and Its Application to LCA and IOA and Related In-dustrial Ecology Tools: Principles and Promise. J. Environ. Acccounting Manag. Volume 3, 151–167. doi:10.5890/JEAM.2015.06.005

Heijungs, R., 2012. Spatial differentiation, GIS-based regionalization, hyperregionalization, and the boundaries of LCA., in: Environment and Energy (Editorial Series of Italian Commodity Sci-ence Academy and Engineering Association of Messina). FrancoAngeli, Milano, Italy, pp. 165–176.

Heijungs, R., 1996. Identification of key issues for further investigation in improving the reliability of life-cycle assessments. J. Clean. Prod. 4, 159–166. doi:10.1016/S0959-6526(96)00042-X

Heijungs, R., Huijbregts, M., 2004. A review of approaches to treat uncertainty in LCA. Complex In-tegr Resour Manag Proc 2nd Bienn Meet Int Env. Model Softw Soc.

Heijungs, R., Kleijn, R., 2001. Numerical approaches towards life cycle interpretation five examples. Int. J. Life Cycle Assess. 6, 141–148. doi:10.1007/BF02978732

Heijungs, R., Suh, S., 2002. The computational structure of life cycle assessment. Kluwer Academic Publishers, Dordrecht.

Heijungs, R., Suh, S., Kleijn, R., 2005. Numerical Approaches to Life Cycle Interpretation - The case of the Ecoinvent'96 database (10 pp). Int. J. Life Cycle Assess. 10, 103–112. doi:10.1065/lca2004.06.161

Hellweg, S., Milà i Canals, L., 2014. Emerging approaches, challenges and opportunities in life cycle assessment. Science 344, 1109–1113.

Helmes, R.J.K., Huijbregts, M.A.J., Henderson, A.D., Jolliet, O., 2012. Spatially explicit fate factors of phosphorous emissions to freshwater at the global scale. Int. J. Life Cycle Assess. 17, 646–654. doi:10.1007/s11367-012-0382-2

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — Global Soil Information Based on Automated Mapping. PLoS ONE 9, e105992. doi:10.1371/journal.pone.0105992

Hevner, A., Chatterjee, S., 2010. Design Research in Information Systems: Theory and Practice. Springer.

Huijbregts, M.J., 1998. Application of uncertainty and variability in LCA. Int. J. Life Cycle Assess. 3, 273–280. doi:10.1007/BF02979835

ISO, 2006. ISO 14044. Environmental management - Life cylce assessment - Requirements and guidelines.

Lenzen, M., 2000. Errors in Conventional and Input-Output—based Life—Cycle Inventories. J. Ind. Ecol. 4, 127–148. doi:10.1162/10881980052541981

Lenzen, M., Dey, C., 2000. Truncation error in embodied energy analyses of basic iron and steel pro-ducts. Energy 25, 577–585. doi:10.1016/S0360-5442(99)00088-2

Lesage, P., Samson, R., 2013. The Quebec Life Cycle Inventory Database Project. Int. J. Life Cycle As-sess. 1–8. doi:10.1007/s11367-013-0593-1

Majeau-Bettez, G., Strømman, A.H., Hertwich, E.G., 2011. Evaluation of Process- and Input–Output-based Life Cycle Inventory Data with Regard to Truncation and Aggregation Issues. Environ. Sci. Technol. 45, 10170–10177. doi:10.1021/es201308x

Monfreda, C., Ramankutty, N., Foley, J.A., 2008. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Glob. Biogeochem. Cycles 22. doi:10.1029/2007GB002947.

Mueller, N.D., Gerber, J.S., Johnston, M., Ray, D.K., Ramankutty, N., Foley, J.A., 2012. Closing yield gaps through nutrient and water management. Nature 490, 254–257. doi:10.1038/nature11420

Mutel, C., 2012. Framework and Tools for Regionlization in Life Cycle Assessment. ETH, Zurich.

Mutel, C.L., Hellweg, S., 2009. Regionalized Life Cycle Assessment: Computational Methodology and Application to Inventory Databases. Environ. Sci. Technol. 43, 5797–5803. doi:10.1021/es803002j

Mutel, C.L., Pfister, S., Hellweg, S., 2012. GIS-based regionalized life cycle assessment: how big is small enough? Methodology and case study of electricity generation. Environ. Sci. Technol. 46, 1096–1103. doi:10.1021/es203117z

Nemecek, T., Bengoa, X., Lansche, J., Mouron, P., Riedener, E., Rossi, V., Humbert, S., 2015. Methodo-logical Guidelines for the Life Cycle Inventory of Agricultural Products. World Food LCA Database (WFLDB), Version 3.0.

Nemecek, T., Schnetzer, J., Reinhard, J., 2014. Updated and harmonised greenhouse gas emissions for crop inventories. Int. J. Life Cycle Assess. 1–18. doi:10.1007/s11367-014-0712-7

Panagos, P., Borelli, P., Meusburger, K., 2015. A New European Slope Length and Steepness Factor (LS-Factor) for Modeling Soil Erosion by Water. Geosciences 5, 117–126.

Panagos, P., Meusburger, K., Ballabio, C., Borrelli, P., Alewell, C., 2014. Soil erodibility in Europe: A high-resolution dataset based on LUCAS. Sci. Total Environ. 479–480, 189–200. doi:10.1016/j.scitotenv.2014.02.010

Peffers, K., Tuunanen, T., Rothenberger, M., Chatterjee, S., 2007. A Design Science Research Methodology for Information Systems Research. J Manage Inf Syst 24, 45–77.

Read L., L., 1958. I, pencil.

Rebitzer, G., Ekvall, T., Frischknecht, R., Hunkeler, D., Norris, G., Rydberg, T., Schmidt, W.-P., Suh, S., Weidema, B.P., Pennington, D.W., 2004. Life cycle assessment: Part 1: Framework, goal and scope definition, inventory analysis, and applications. Environ. Int. 30, 701–720. doi:10.1016/j.envint.2003.11.005

Reinhard, J., Emmenegger, M.F., Widok, A.H., Wohlgemuth, V., 2011. RSB tool: a LCA tool for the as-sessment of biofuels sustainability, in: Proceedings of the Winter Simulation Conference. Winter Simulation Conference, Phoenix, Arizona, pp. 1048–1059.

Rørbech, J.T., Vadenbo, C., Hellweg, S., Astrup, T.F., 2014. Impact Assessment of Abiotic Resources in LCA: Quantitative Comparison of Selected Characterization Models. Environ. Sci. Technol. 48, 11072–11081. doi:10.1021/es5023976

Scherer, L., Pfister, S., 2015. Modelling spatially explicit impacts from phosphorus emissions in agriculture. Int. J. Life Cycle Assess. 20, 785–795. doi:10.1007/s11367-015-0880-0

Seto, K.C., Reenberg, A., Boone, C.G., Fragkias, M., Haase, D., Langanke, T., Marcotullio, P., Munroe, D.K., Olah, B., Simon, D., 2012. Urban land teleconnections and sustainability. Proc. Natl. Acad. Sci.

Sonnemann, G., Vigon, B. (eds)., 2011. Global Guidance Principles for Life Cycle Assessment Data-bases: A Basis for Greener Processes and Products (Shonan Guidance Principles).

Steubing, B., Wernet, G., Reinhard, J., Bauer, C., Moreno-Ruiz, E., 2016. The ecoinvent database versi-on 3 (part II): analyzing LCA results and comparison to version 2. Int. J. Life Cycle Assess. 21, 1269–1281. doi:10.1007/s11367-016-1109-6

Suh, S., Huppes, G., 2005. Methods for Life Cycle Inventory of a product. J. Clean. Prod. 13, 687–697. doi:10.1016/j.jclepro.2003.04.001

Suh, S., Huppes, G., 2002. Missing inventory estimation tool using extended input-output analysis. Int. J. Life Cycle Assess. 7, 134–140. doi:10.1007/BF02994047

Tillman, A.-M., 2000. Significance of decision-making for LCA methodology. Environ. Impact Assess. Rev. 20, 113–123. doi:10.1016/S0195-9255(99)00035-9

USGS, 2016. Mineral Resources Data System (MRDS) [WWW Document]. URL https://mrdata.usgs.gov/mrds/

Weidema, B., Bauer, C., Hischier, R., Mutel, C., Nemecek, T., Reinhard, J., Vadenbo, C., Wernet, G., 2013. Overview and methodology. Data quality guideline for the ecoinvent database version 3.

Werner, F., 2002. Interdependencies between LC-Modelling and the use of LCA in product design-related decision situations. With special emphasis on the influence of cognitive models and values on the modelling of reuse & recycling and other end-of-life options (Diss. ETH Nr. 14’750). Zurich.

Wernet, G., 2012. ecoinvent v3 - Internationalization of the data supply and international collaborations. Presentation at DF48, Dübendorf.

Zah, R., Faist, M., Reinhard, J., Birchmeier, D., 2009. Standardized and simplified life-cycle assessment (LCA) as a driver for more sustainable biofuels. J. Clean. Prod. 17, Supplement 1, S102–S105. doi:10.1016/j.jclepro.2009.04.004

Zah, R., Gmuender, S., Gauch, M., 2012. Evaluación del ciclo de vida de la cadena de producción de biocombustibles en Colombia (No. Estrategias de energía sostenible y biocombustibles para Colombia ATN/JC-10826-CO y ATN/JF-10827-CO). Banco Interamericano de Desarrollo (BID).

PART II – PUBLICATIONS THAT CONSTITUTE THIS DISSERTATION

2 Article I: Contribution-based prioritization of LCI database improvements. Method design, demonstration and evaluation

Juergen Reinhard°*, Christopher Mutel◊, Gregor Wernet#, Rainer Zah¥, Lorenz M. Hilty°*,

°*University of Zurich, Department of Informatics, Informatics and Sustainability Research Group, Binzmuehlestrasse 14, 8050 Zurich, Switzerland.*

◊ *Laboratory for Energy Systems Analysis, Paul Scherrer Institute, 5232 Villingen PSI, Switzerland*

#*Swiss Centre for Life Cycle Inventories, Ecoinvent, Lerchenfeldstrasse 5, 9014 St. Gallen, Switzerland*

¥*Quantis Serl, Reitergasse 11, 8600 Zurich, Switzerland*

**Empa Materials Science and Technology, Technology and Society Laboratory, Ueberlandstrasse 129, 8600 Duebendorf, Switzerland. Tel. +41 58 765 4207*

Status: *Published as "Contribution-based prioritization of LCI database improvements: Method design, demonstration, and evaluation." in Environmental Modelling & Software 86, 204–218. doi:10.1016/j.envsoft.2016.09.018*

Supplementary material

This article contains supplementary material (section 2.9) which complements the article with additional figures and discussions.

Abstract

The largest Life Cycle Inventory (LCI) databases contain about 10,000 unit process datasets. Assuming that updating one dataset would, on average, require one working day, then updating the entire database would roughly require the continuous work of five workers for ten years. Methods are therefore needed to prioritize datasets to be updated, and such a method should be able to identify the unit processes that contribute the most across different Life Cycle Impact Assessment (LCIA) methods. To date, such methods are not available.

This paper presents an operational prioritization method. We demonstrate and evaluate our method by applying it to the ecoinvent database. The case study shows that prioritization of improvement efforts is highly useful because a robust nucleus of unit processes proves to be consistently important across all product systems and many LCIA methods. Focusing research efforts on these processes allows the effective improvement of the LCI database.

Keywords

Prioritization; Life Cycle Inventory database; Meta analysis; Contribution analysis; Life Cycle Assessment

Software and data availability

All modelling in this paper is done in MATLAB®; the code and data for the contrived example can be forwarded by the first author upon request.

2.1 Introduction

2.1.1 LCA: Background data is key but uncertain

In our modern economy, the provision, use, and disposal of almost any product involves complex globalized networks—called product systems—consisting of thousands of interlinked human activities (Hellweg and Milà i Canals, 2014). Accounting for the cumulated environmental impact of such product systems requires tracing and measuring the metabolism of all activities associated with a particular product. Life Cycle Assessment (LCA) is one of the most prominent techniques developed for this purpose (ISO, 2006). It focuses on the "compilation and environmental evaluation of all processes, outputs and the potential environmental impacts associated with a product system" (ISO, 2006, p. 2) with the goal of pinpointing ecological weaknesses, comparing alternatives, evaluating the main environmental impacts, designing new products, measuring the environmental relevance of a material or product, and establishing recommendations for actions (Guinée et al., 2002).

The LCA procedure is structured around a static model viewing the product system as a network of basic building blocks called unit processes. A unit process represents one specific activity or a group of activities in the product system and records (i) the *intermediate exchanges* from and to the technosphere, i.e., the input of usable energy and raw materials and the output of products and waste and (ii) the *exchanges with environment*, i.e., the input of natural resources and output of emissions (Finnveden et al., 2009). The unit processes are linearly linked by means of their intermediate exchanges. The calculation of the product system for the reference flow of interest results in the sum of all *exchanges with the environment* for the entire life cycle, called Life Cycle Inventory (LCI). Life Cycle Impact Assessment (LCIA)—the next step within the LCA procedure—then takes the inventory data on these exchanges as an input to determine the impacts on the environment.

Although this procedure reduces the complexity of the real-world system under study, LCA remains a very data-intensive technique. A typical product life cycle covers thousands of unit processes, each of which needs to be described with exchange flow data. This information can usually not be gathered within a specific project due to the high cost of data collection. It is therefore common practice to focus data collection efforts on selected activities that reflect the space for action—these activities are together called the foreground system—and to use generic data from Life Cycle Inventory (LCI) databases to model the remaining activities, called the background system (Bourgault et al., 2012; Tillman, 2000). The background system usually covers up to 99% of the unit processes in the product system; only in rare cases does the number of unit processes modeled explicitly in the foreground system exceed 5%. Bearing this in mind, background or Life Cycle Inventory (LCI) databases can be considered the backbone of any LCA study. They provide the dominant share of the building blocks required for any LCA: aggregated and/or disaggregated unit process data. Therefore, the available quantity and quality of unit process data provided by LCI databases are of utmost importance.

However, the unit process data provided by LCI databases are characterized by a high degree of uncertainty (Groen et al., 2014). This has several reasons. First, it is not feasible to match each original activity in the real world with a description of a unit process in the LCI database. Consequently, the exchange flow values recorded in a unit process often represent assumed average conditions of a whole country or larger region across a given time period and across different instances of real processes. However, industrial production characteristics and environmental conditions can vary significantly over geographic space, time, and process instances (Huijbregts, 1998). Second, the quantitative data required to compile an accurate representation of a specific activity may be unavailable, wrong, or unreliable (Ciroth et al., 2004; Heijungs and Huijbregts, 2004). Third, the growth of LCI databases during the last twenty years was primarily supported by a local funding structure. Consequently, LCI databases are biased in their coverage of real-world activities towards their local origin, whereas supply chains are global. Although current developments improve on this shortcoming (see for example Wernet (2012)), today's LCI databases are far from being complete, even on the level of generic data.

Heijungs and Huijbregts (2004) present various approaches to address these uncertainties;

- the statistical approach (applying methods from statistics to existing LCI data),
- the constructivist approach (involving stakeholders),
- the legal approach ("truth" decreed by authoritative bodies), and
- the scientific approach (doing more research to create or improve primary LCI data).

While all approaches have merit, this paper explores a combination of the statistical approach with the scientific approach: Statistical methods provide the focus for the scientific approach in the systematic improvement of quality in LCI databases.

2.1.2 Background databases: Effective prioritization is challenging

We believe that a statistical approach has particular merit because effectively prioritizing improvement efforts is becoming increasingly difficult due to the growing numbers of unit processes stored in existing databases. For example, depending on the system model[26], the current version of ecoinvent (3.1) includes

[26] A system model provides a set of rules that specify how activity datasets are linked to form product systems. Ecoinvent version 3.1 offers three system models to choose from. They differ mainly with regard to their handling of

between 10,305 and 11,332 unit processes. If updating one unit process would, on average, require one worker day[27], a systematic update of the entire database within one year would require the continuous work of more than 50 people. Capacities for such extensive improvement efforts are typically not available. In addition, the desired quality and quantity of databases is a "constantly moving target." It is the "continuing evolution in consumer preferences, market and industry imperatives, and public policy which forces continuous development and improvement of datasets and methodologies for LCA to meet these needs" (Sonnemann and Vigon, 2011, p. 98). Compliance with these developments typically ties up a lot of the limited workforce. Consequently, improvements in data quality and quantity are so far rather directed by external data availability and situation-driven requirements than by systematic choice. To our knowledge, none of the LCI databases applies statistical methods to prioritize LCI database improvements.

Moreover, LCI databases represent highly interconnected systems, where almost every unit process is involved in every product system. As noted by Heijungs (2012, p. 172), "due to the interlinked nature of many processes, seemingly alien[28] parts of the database are involved in all LCAs". This makes it difficult to align improvement efforts effectively to the unit processes that matter most or to anticipate the changes induced by an update in advance. For example, each of the roughly 4,000 product systems in ecoinvent v.2.2 require between 2,100 and 2,300 processes (Heijungs, 2012). That is, the system-wide representation of almost every product system involves more than half of the unit processes in the entire database (Heijungs, 2012). Vice versa, updating an existing or integrating a new unit process can affect the results for a large part of the products. Identifying the unit processes that matter most would allow untangling these complex interdependencies and facilitating the effective and efficient improvement of the entire LCI database.

---

the multi-functionality problem (system expansion vs. allocation), the use of average or marginal suppliers, and their assessment of by-product treatments (Weidema et al., 2013).

[27] The workload for updating a datasets varies greatly. Depending on the given level of completeness of a dataset, it might take weeks (e.g. when a global datasets is disaggregated into many country specific datasets) or just an hour (e.g., when the quantity of one emission flow in one dataset is updated). Updating a dataset typically involves data collection, data entry/manipulation and data submission to peer review (using a software tool called EcoEditor). Any change in an existing datasets or the submission of new datasets requires a peer review to be accepted into the database (Weidema et al., 2013). Therefore, one worker day should be considered as a rough but realistic estimate of the average effort associated with updating a datasets.

[28] For example, the product system associated with the provision of one kWh of high voltage electricity (produced by a hydro power plant in an alpine Swiss region) involves "alien" processes such like airport infrastructure, sugar (from sugar cane), kraft paper, and many more.

2.1.3 Prioritization of LCI database improvements

In general, we can distinguish two perspectives to prioritize of LCI database improvements; external and internal prioritization. External prioritization aligns inventory efforts to importance based on an external reference such as environmentally extended input-output (EEIO) databases. For example, Majeau-Bettez et al. (2011) contrasted the economic, environmental, and structural importance of the economic sectors (obtained from an EEIO database) with the inventory effort (the number of corresponding unit processes) in version 2.1 of the ecoinvent database in order to identify the sectors that are most underrepresented. The discrepancy between the sectorial importance in the EEIO database and the corresponding inventory effort leads them to suspect a "suboptimal allocation of LCA inventory resources" (Majeau-Bettez et al., 2011, p. 10175).

Internal prioritization aligns inventory efforts by identifying key environmental issues in the existing data. To the best of our knowledge, LCI database improvement efforts have not been systematically guided by the identification of key elements so far, even though the identification of such elements in LCA studies is well developed (Heijungs, 1996; Heijungs et al., 2005; Heijungs and Kleijn, 2001). Already in 1996, Heijungs (1996) published a method to identify data elements that are characterized by both high uncertainty and a high contribution. Key issue analysis is the procedure that assesses a result in terms of the uncertainties and contributions of its constituting data elements to identify the elements whose improvement would have the greatest effect (see Figure 2-1). The purpose of the uncertainty dimension is to identify the data elements which dominate the propagation of uncertainty (Heijungs, 1996), that of the contribution dimension to point out those data elements that make the highest contribution (Heijungs and Kleijn, 2001).

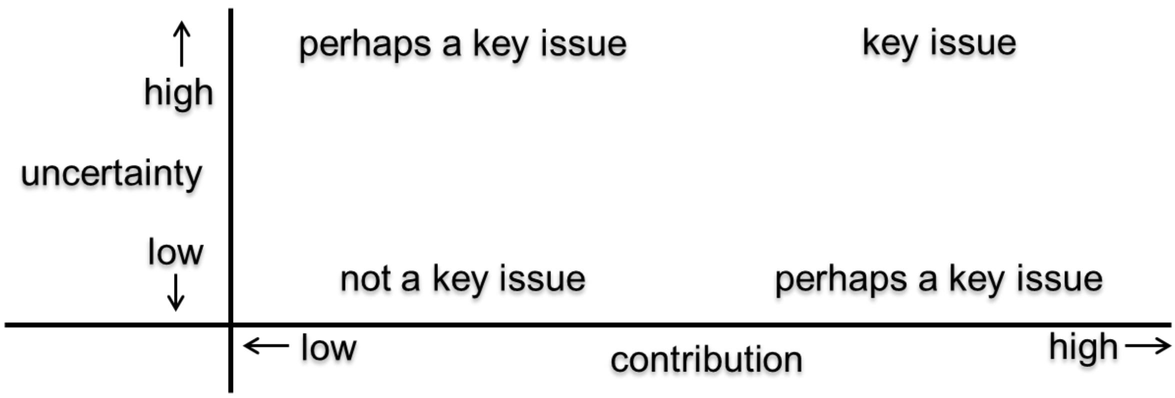


Figure 2-1: A key issue is a data element that is characterized by high uncertainty and high contribution. Data elements with a high score on one dimension and a low score on the other dimension should be analyzed in further detail. Data elements with low contribution and low uncertainty are not key issues. Source: (Heijungs, 1996)

This paper focuses on the contribution dimension. We describe a method for database-wide contribution analysis (CA) and corresponding indicators that facilitate the internal prioritization of improvement efforts based on the identification of key processes, i.e., unit processes with a consistently large relative contribution.

2.1.4 Contribution-based prioritization

2.1.4.1 Contribution analysis (CA)

In LCA, CA is used for the identification of key elements in the life cycle, i.e., the importance of life cycle stages, unit processes, or elementary flows involved. CA is quite common in LCA and implemented in all commercially available software tools (Ciroth, 2013; Ebner, 2013; Goedkoop and Oele, 2004). To date, however, CA has not been formalized under the aspect of prioritizing LCI database improvements, even though some studies use CA for the analysis of some specific features of LCI databases. For example, Frischknecht et al. (2007) analyzed the relative environmental contribution of capital goods throughout the ecoinvent database for numerous LCIA indicators to decide whether capital goods can be generally excluded from LCA applications. Lesage and Samson (2013) used CA to identify datasets for recontextualization. They regionalized[29] the direct electricity consumption of all datasets in version 2.2 of the ecoinvent database "for which direct electricity consumption contributed at least 25% of the life cycle impacts" (Lesage and Samson, 2013, p. 5) in order to improve data representativeness for Canada. Rørbech et al. (2014) uses CA to evaluate the relative importance of elementary flows across LCIA resource depletion indicators for a selected sample of version 3.0 of the ecoinvent database in order to identify differences and commonalities in the foci of LCIA indicators.

Our elements of interest are the *relative* contributions of each unit process throughout each product system represented with respect to the full spectrum of LCIA[30] indicators. Such relative contributions can be calculated for two perspectives; the *causer perspective* and the *connector* or *network perspective*. These perspectives represent complementary lenses for our goal of identifying important processes throughout the database.

- In the *causer perspective*, we focus exclusively on the *causing elements* of each unit process. Causing elements are the direct exchanges with the environment that cause environmental impacts. For example, processing oil into electricity causes the emission of 6 kg CO_2 and 1 kg CH_4 per kWh of electricity produced (see matrix **B**, column one in Figure 2-2). This perspective helps to pinpoint the processes with consistently large contributions in terms of environmental interventions.

[29] That is, they replaced the former electricity mix with the Canadian electricity mix of relevance.

[30] In principle, CA can also be applied on the inventory level (Heijungs and Kleijn, 2001). We focus on environmental impacts, since we believe that a CA on the inventory level is of little practical interest in the context of prioritization of improvement efforts.

- In the *connector perspective*, we focus exclusively on the *connecting elements* of each unit process, that is, the intermediate flows from other processes of the technosphere. From this perspective, the environmental impact of one kg of steel is determined by the environmental impact associated with the intermediate flows of 0.12 kWh of electricity and 1.20 kg of iron ore (see matrix **A**, column three in Figure 2-2). That is, we measure the environmental impact not at its actual source, but at the point of transmission. This perspective helps us to pinpoint the unit processes that consistently *link* to large upstream contributions. Such processes represent sensitive *hubs* whose modification can change the results for the overall database considerably.

Realizing a database-wide CA for these perspectives requires the processing of a vast amount of data. For version 3.1 of the ecoinvent database, a CA for one LCIA indicator and one perspective requires the computation of the system-wide contributions of more than 11,000 product systems. There is no default solution for this task. Basically, two kinds of algorithms are available; those based on a sequential procedure and those based on matrix inversion (Bourgault et al., 2012; Heijungs and Suh, 2002, p. 101). The sequential approach is not applicable for our purpose because it either entails high computational costs or truncation errors and – more important – cannot deal with recursion[31], which is often found in product systems (Bourgault et al., 2012; Heijungs and Suh, 2002, p. 101). Algorithms based on matrix inversion can more efficiently be computed and can deal with recursion. However, the standard implementation of matrix inversion (Heijungs and Suh, 2002, p. 168) focuses on one product system and does not allow for process-specific CA as the contributions of each process (our elements of interest) are automatically summed to a total impact score (Bourgault et al., 2012; Mutel and Hellweg, 2009). In order to implement a CA for the causer perspective therefore requires adapting the LCA standard calculation procedure. An implementation for the connector perspective requires the relocation of the direct environmental impacts according to a new procedure.

2.1.4.2 Summary measures

Summary measures "condense a certain aspect of a large set of numbers into one or perhaps a few items" (Heijungs and Suh, 2002, p. 152). We use standard statistical summary measures for investigating key processes with constantly large contributions in the relative contribution matrixes (obtained from the CA), i.e. the unit processes which contribute the most to a result calculated with a given LCIA method. We also use advanced summary measures such as the Gini coefficient or the Concentration ratio. Such measures indicate the utility of prioritization because they summarize the inequality in process importance. In particular, high inequality in process importance indicates high utility of prioritization.

[31] Recursion can "occur if processes require each other as inputs" (Mutel and Hellweg, 2009).

2.1.5 Structure

Our desired end is the prioritization of improvement efforts via the identification of unit processes of systemic importance. Towards this end, this paper covers the following steps:

1. We demonstrate our method using a contrived LCI database example in a formalized but general manner. This shows the key components of our method in their exemplary application and facilitates replication.
2. We demonstrate our method in its application to version 3.1 of the ecoinvent database for two selected LCIA methods (proof of concept).
3. We evaluate our method based on the experience gained in the proof-of-concept application, providing a framework to reason about the general utility of prioritization in LCI databases.

2.2 Materials and method

2.2.1 Contrived example (CE)

We describe our method on the basis of a contrived LCI database example. Figure 2-2 shows the elements of our simplified database consisting of six product systems, their environmental interventions, and the characterization matrix associated with these environmental interventions.

The technosphere matrix, **A**[32]**,** (Figure 2-2, top) expresses a system of electricity, wood, steel, oil, coal, and fuel in matrix notation. *Inputs* and *outputs* are listed by row, *processes* by column. Positive numbers represent the production of outputs, negative numbers the consumption of direct inputs. Each process requires inputs, e.g., oil is needed for electricity production and steel is needed to build the machinery that extracts the oil, and produces outputs, e.g. electricity. **B**, the biosphere matrix, defines the environmental interventions (resources consumed and emissions released) per unit of process (Figure 2-2, middle). In this simple example, we focus exclusively on the emission of carbon dioxide (CO_2) and methane (CH_4). The characterization matrix **W** represents the characterization factors associated with these environmental interventions for two LCIA indicators: climate change (CC) and photochemical oxidant formation (POX) (Figure 2-2, bottom). All of the following calculation procedures will be demonstrated with the LCIA indicator climate change (i.e., the first row of the characterization matrix), which we denote with *w*[33].

[32] Bold capital letters indicate matrixes.

[33] Italic lower-case letters indicate vectors.

			Unit	Electricity generation process	Iron ore production process	Steel production process	Oil production process	Coal extraction process	Fuel production process
A	Products	Electricity	kWh	1.00	-	-0.12	-	-0.56	-0.02
		Iron ore	kg	-	1.00	-1.20	-	-	-
		Steel	kg	-	-0.02	1.00	-0.02	-0.12	-
		Oil	kg	-0.50	-	-	1.00	-	-1.40
		Coal	kg	-	-	-	-	1.00	-
		Fuel	kg	-	-0.60	-	-0.04	-	1.00
B	elem. flows	CO_2	kg	6.00	0.40	-	1.00	2.00	1.00
		CH_4	kg	1.00	0.02	-	0.05	0.10	0.10

	LCIA		Unit	CO_2 (kg)	CH_4 (kg)
W	CC	CO_2-eq.	kg	1.00	28.00
	POX	NMVOC	kg	-	0.0101

Figure 2-2: Technosphere matrix A (top), biosphere matrix B (middle), and characterization matrix W of our contrived LCI database example. '-' denotes zero.

2.2.2 Calculating relative contribution matrixes

2.2.2.1 Causer perspective

Figure 2-3 shows the workflow for the computation of the relative contribution matrix for the causer perspective. The workflow has to be repeatedly executed for every LCIA indicator of interest, i.e. for each row, k, of the characterization matrix W (which we denote with w).

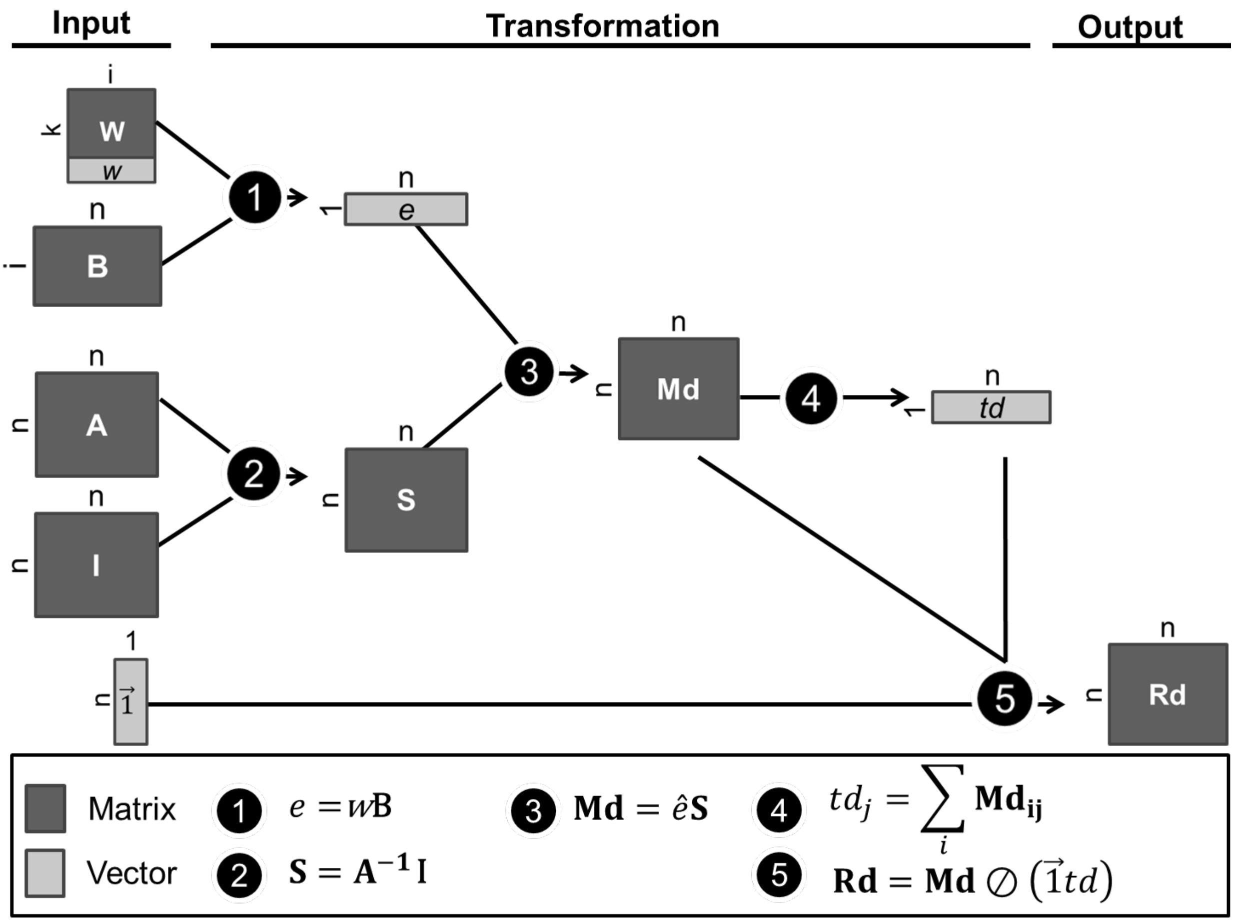


Figure 2-3: Workflow to compute the direct relative contribution of each unit process throughout all product systems for the causer perspective. The inputs required are the identity matrix I (of size $n \times n$), the technosphere matrix A, the

biosphere matrix B, one row from the characterization matrix W, (which we denote as w) and a vector $\vec{1}$, which represents a vector of ones of size $n \times 1$. All transformations are explained on the basis of the CE.

This workflow is applicable to any LCI database in the form of disaggregated technosphere matrix **A** and a biosphere matrix **B.** Moreover, a corresponding characterization matrix **W** is needed. The implementation into our method is largely based on the LCA standard calculation procedure. The key adaptations for our purpose concerns:

- the diagonalization (expressed by the hat, ^) of vector, e in equation 3 which facilitates an in-depth analysis of the contribution per process as it allows the process-specific analysis of the typically aggregated LCIA results.
- the use of a demand matrix instead of a demand vector. In order to compute the entire LCI database simultaneously, we multiply the inverted technosphere matrix, **A**, with the identity matrix, **I** (see equation 2) which specifies a unitary product output for each product system in the database (Heijungs and Suh, 2002, p. 84).
- the avoidance of computational expensive matrix manipulations in order to ensure its applicability to large matrixes.

In the following, we will walk through each step of this workflow on the basis of the CE. First, we calculate the direct environmental impacts associated with the use of one unit of a unit process, e, by multiplying the characterization vector for climate change, w, with the biosphere matrix (equation 1).

$$e = w\mathbf{B} \tag{1}$$

The vector e, then represents for each unit of a unit process the environmental impacts directly associated with its elementary flows (Figure 2-4 , top).

				Processes					
			Unit	Electricity generation process	Iron ore production process	Steel production process	Oil production process	Coal extraction process	Fuel production process
e		**Direct impacts**	**kg CO_2 eq.**	**34.00**	**0.96**	**-**	**2.40**	**4.80**	**3.80**
I	Products	Electricity	kWh	1.00	-	-	-	-	-
		Iron ore	kg	-	1.00	-	-	-	-
		Steel	kg	-	-	1.00	-	-	-
		Oil	kg	-	-	-	1.00	-	-
		Coal	kg	-	-	-	-	1.00	-
		Fuel	kg	-	-	-	-	-	1.00
tn^T		**Cumulated impacts**	**kg CO_2 eq.**	**35.43**	**6.12**	**11.40**	**2.75**	**26.11**	**8.02**
p		**Upstream proportion**		**0.04**	**0.85**	**1.00**	**0.20**	**0.82**	**0.56**

Figure 2-4: Vector e, the direct environmental impacts per process, the identify matrix I (representing a unitary demand for each product system), vector tnT representing the cumulated (direct and upstream) environmental impact per unit of process and p, the proportion of process specific impacts which are caused upstream. '-' denotes zero.

We next calculate the amount needed of each process to satisfy a unitary demand across all available product systems—the supply matrix **S**—by computing the inverse of the technosphere matrix **A** and multiplying it by the identity matrix **I** (equation 2). The identity matrix represents a unitary product demand (reference flow) for all product system (Heijungs and Suh, 2002, p. 84) (see Figure 2-4, middle).

$$\mathbf{S} = \mathbf{A}^{-1}\mathbf{I} \tag{2}$$

Figure 2-5 (top) shows **S** for our contrived example. The circular characteristic of the system causes that the demand for one unit of a product requires more than one unit of the corresponding process, e.g., 1.05 units of the iron ore production process are required for the production of 1 kg of iron ore.

Transforming the sum vector *e* into a diagonal matrix and multiplying it by the supply matrix **S** delivers the direct environmental impacts associated with every single process specified in **S** throughout all product systems. That is, each column of this matrix represents the system-wide, process-specific environmental impacts for supplying one unit of the product referred to by that column.

$$\mathbf{Md} = \hat{e}\mathbf{S} \tag{3}$$

$$td_j = \textstyle\sum_i \mathbf{Md}_{\mathrm{ij}} \tag{4}$$

Consequently, the sum vector of **Md**, *td*, then expresses the total environmental impact associated with the production of a unitary demand of each product system in the database (Figure 2-5, middle).

				Product systems					
			Unit	Electricity 1 kWh	Iron ore 1 kg	Steel 1 kg	Oil 1 kg	Coal 1 kg	Fuel 1 kg
S	Processes	Electricity generation	process	**1.00**	0.02	0.14	0.00	0.58	0.03
		Iron ore production	process	0.01	**1.05**	1.26	0.03	0.16	0.04
		Steel production	process	0.01	0.04	**1.05**	0.02	0.13	0.03
		Oil production	process	0.54	0.94	1.20	**1.09**	0.45	1.53
		Coal extraction	process	-	-	-	-	**1.00**	-
		Fuel production	process	0.03	0.67	0.80	0.06	0.11	**1.08**
Md	Processes	Electricity generation	kg CO_2 eq.	34.07	0.62	4.83	0.13	19.66	0.86
		Iron ore production	kg CO_2 eq.	0.01	1.01	1.21	0.03	0.15	0.04
		Steel production	kg CO_2 eq.	-	-	-	-	-	-
		Oil production	kg CO_2 eq.	1.30	2.26	2.87	2.60	1.07	3.67
		Coal extraction	kg CO_2 eq.	-	-	-	-	4.80	-
		Fuel production	kg CO_2 eq.	0.11	2.53	3.05	0.23	0.43	4.12
td		**Total impacts**	**kg CO_2 eq.**	**35.49**	**6.41**	**11.96**	**2.99**	**26.11**	**8.69**

				Product systems						Summary Measures		
			Unit	Electricity relative contribution	Iron ore relative contribution	Steel relative contribution	Oil relative contribution	Coal relative contribution	Fuel relative contribution	*Ø*	*me*	*sd*
Rd	Processes	Electricity generation		0.96	0.10	0.40	0.04	0.75	0.10	0.39	0.25	0.35
		Iron ore production		0.00	0.16	0.10	0.01	0.01	0.00	0.05	0.01	0.06
		Steel production		-	-	-	-	-	-	-	-	-
		Oil production		0.04	0.35	0.24	0.87	0.04	0.42	0.33	0.30	0.28
		Coal extraction		-	-	-	-	0.18	-	0.03	-	0.07
		Fuel production		0.00	0.39	0.26	0.08	0.02	0.47	0.20	0.17	0.18
		Sum		**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.72**	

Figure 2-5: The supply matrix S (top), the direct contribution matrix Md (middle), and the relative contribution matrix Rd (bottom) for the CE. '-' denotes zero. Ø stands for the arithmetic mean, me for median, sd for standard deviation.

Finally, we compute, for each product system, the relative contribution of each process in relation to the contribution of all processes. In order to allow efficient computation, we expand vector td $(1\times n)$ to a matrix that matches the dimension of $\mathbf{Md}$, i.e. $(n \times n)$. We perform this expansion[34] by multiplying td with $\vec{1}$, which represents a vector of ones of size $(n \times 1)$.The element-by-element division ($\oslash$) of $\mathbf{Md}$ and the expanded vector td results in the relative environmental impact contribution matrix $\mathbf{Rd}$[35] (Figure 2-5, bottom).

$$\mathbf{Rd} = \mathbf{Md} \oslash \left(\vec{1}td\right) \tag{5}$$

$\mathbf{Rd}$ shows the relative contribution of each process throughout each product system in the database. To put it another way, it tells us for each process-product system combination the relative loss if we remove all elementary flows of the unit process. Note that $\mathbf{Rd}$ only shows a relative contribution if the process includes elementary flows of relevance for the LCIA indicator used. For example, since steel production causes no direct emissions, its total and relative contributions amount to zero.

2.2.2.2 Connector perspective

Figure 2-6 shows the workflow for the computation of the relative contribution matrix of the connector perspective. Note that almost all of the required input data represents intermediate data from the causer perspective.

[34] In Matlab, the binary singleton expansion function (bsxfun) allows performing this expansion automatically.

[35] This operation is equal to $\mathbf{Md} * \widehat{td}^{-1}$.

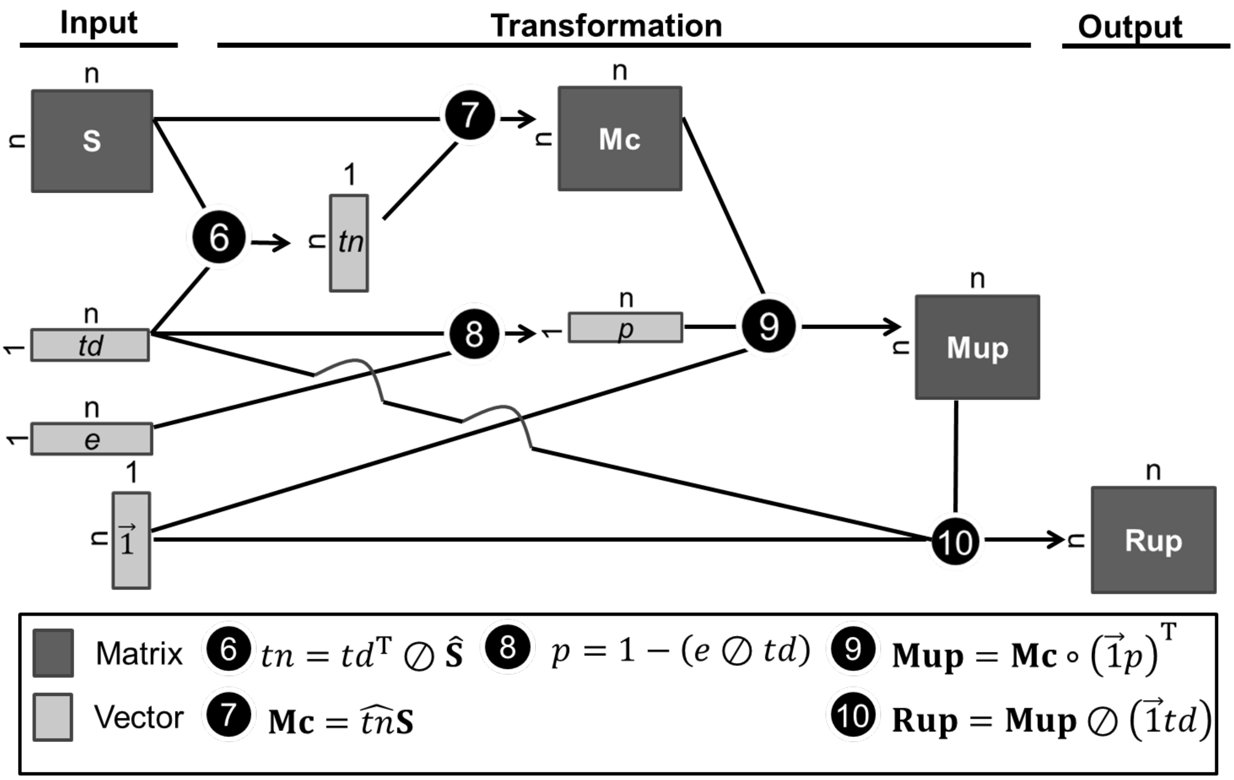


Figure 2-6: Workflow to compute the direct relative contribution of each unit process throughout all product systems for the connector perspective. The inputs required are the supply matrix S, the total environmental impact per product system, td, the direct environmental impacts associated with the use of one unit of a unit process, e, and vector $\vec{1}$, which represents a vector of ones of size $n \times 1$. All transformations are explained on the basis of the contrived example (CE).

For the connector's perspective, we relocate the environmental impact obtained in the causer perspective (i.e., in terms of *td*) as follows. First, we scale the transposed total environmental impact of each product system, *td*, by means of the diagonal values of the supply matrix **S** in order to receive *tn*, the cumulated (direct and upstream) environmental impacts associated with the use of one unit of each process (see Figure 2-4, middle). When we compute $\mathbf{A}^{-1}$, the circular characteristic of the system causes that the demand for one unit of a product requires more than one unit of the corresponding process, e.g. the demand for 1 kg of fuel requires 1.08 fuel production processes. The mentioned re-scaling ensures that the environmental impacts refer to exactly one unit of a process. For example, by dividing the total environmental impacts associated with the production of 1 kg of fuel (8.69 kg CO_2-eq.) by the corresponding amount of fuel production processes (1.08 processes, see Figure 2-5) we obtain the cumulated impact per fuel production process (8.02[36] kg CO_2-eq., see Figure 2-4).

$$tn = td^{\mathrm{T}} \oslash \hat{\mathbf{S}} \tag{6}$$

$$\mathbf{Mc} = \widehat{tn}\mathbf{S} \tag{7}$$

[36] Difference results from rounding.

Next, we diagonalize tn and multiply it by **S**. This operation multiplies the cumulated impact per process by its corresponding process demand in the supply matrix **S** in order to calculate **Mc,** the cumulated (direct and upstream) environmental impacts associated with the amount of processes needed to satisfy a unitary product demand for all product systems (Figure 2-7, top). For example, multiplying the cumulate impact of the fuel production process (8.02 kg CO_2-eq.) with the corresponding demand of fuel production processes across the product systems recorded in **S** (1 kWh electricity demands 0.03 fuel production processes, 1 kg iron ore demands 0.67 fuel production processes, etc.) yields the cumulated impacts of each fuel production process scaled to the respective process need indicated by the supply matrix **S** (0.24 kg CO_2-eq. for electricity, 5.34 kg CO_2-eq. for iron ore, etc.). Therefore, the diagonal values of **Mc** are equal to td.

We next calculate the upstream proportion for each process, p, as the deviation between one and the element-by-element ratio of the direct process contributions, e, and the total environmental impacts, td (Figure 2-4, bottom).

$$p = 1 - (e \oslash td) \tag{8}$$

$$\mathbf{Mup} = \mathbf{Mc} \circ \left(\vec{1}p\right)^{\mathrm{T}} \tag{9}$$

We obtain **Mup** the environmental impact associated exclusively with the upstream impacts (Figure 2-7, middle), by calculating the Hadamard product (∘) of the cumulated impact of each process in **Mc** and the proportion caused by its upstream contributions. In order to allow that each process impact in **Mc** is multiplied with its corresponding upstream proportion, we need to expand vector p $(n \times 1)$ to match the dimensions of **Mc**, i.e. $(n \times n)$. We perform this expansion by multiplying vector p with $\vec{1}$, which represents a vector of ones of size $(k \times 1)$, and transposing the result.

Finally, we apply element-by-element division to the expanded sum vector td and the matrix **Mup**, to obtain the relative contribution matrix for the upstream impacts **Rup** (Figure 2-7, bottom).

$$\mathbf{Rup} = \mathbf{Mup} \oslash \left(\vec{1}td\right) \tag{10}$$

It is noteworthy that we divide **Mup** by td, and not tup, the sum of the upstream contributions of all processes per product system. Therefore, each column-row combination in **Rup** expresses the upstream contribution of a particular process in relation to the total environmental impact of the corresponding product system.

			Product systems					
			Electricity	Iron ore	Steel	Oil	Coal	Fuel
		Unit	1 kWh	1 kg	1 kg	1 kg	1 kg	1 kg
Mc	**Processes**							
	Electricity generation	kg CO_2 eq.	**35.49**	0.64	5.03	0.14	20.48	0.90
	Iron ore production	kg CO_2 eq.	0.08	**6.41**	7.71	0.16	0.97	0.23
	Steel production	kg CO_2 eq.	0.13	0.45	**11.96**	0.25	1.51	0.36
	Oil production	kg CO_2 eq.	1.49	2.59	3.29	**2.99**	1.23	4.21
	Coal extraction	kg CO_2 eq.	-	-	-	-	**26.11**	-
	Fuel production	kg CO_2 eq.	0.24	5.34	6.44	0.48	0.91	**8.69**
Mup	**Processes**							
	Electricity generation	kg CO_2 eq.	1.49	0.03	0.21	0.01	0.86	0.04
	Iron ore production	kg CO_2 eq.	0.07	5.45	6.55	0.14	0.83	0.20
	Steel production	kg CO_2 eq.	0.13	0.45	11.96	0.25	1.51	0.36
	Oil production	kg CO_2 eq.	0.29	0.51	0.65	0.59	0.24	0.83
	Coal extraction	kg CO_2 eq.	-	-	-	-	21.31	-
	Fuel production	kg CO_2 eq.	0.13	1.53	3.62	0.27	0.51	4.89
tup	**Sum**	**kg CO_2 eq.**	**2.12**	**7.97**	**22.99**	**1.25**	**25.26**	**6.31**

			Product systems						Summary Measures		
			Electricity	Iron ore	Steel	Oil	Coal	Fuel	*Ø*	*me*	*sd*
		Unit	relative contribution	relative contribution	relative contribution	relative contribution	relative contribution	relative contribution			
Rup	**Processes**										
	Electricity generation		0.04	0.00	0.02	0.00	0.03	0.00	0.02	0.01	0.02
	Iron ore production		0.00	0.85	0.55	0.05	0.03	0.02	0.25	0.04	0.33
	Steel production		0.00	0.07	1.00	0.08	0.06	0.04	0.21	0.06	0.35
	Oil production		0.01	0.08	0.05	0.20	0.01	0.10	0.07	0.07	0.06
	Coal extraction		-	-	-	-	0.82	-	0.14	-	0.30
	Fuel production		0.00	0.47	0.30	0.09	0.02	0.56	0.24	0.20	0.22
	Sum		**0.06**	**1.47**	**1.92**	**0.42**	**0.97**	**0.73**	**0.93**	**0.38**	

Figure 2-7: The environmental impact matrix Mc (top), the total and relative upstream contribution matrixes Mup (middle), and Rup (bottom) for our CE. Rup is obtained by dividing each element in Mup by td (Figure 2-5). '-' denotes zero. Ø stands for the arithmetic mean, me for median, sd for standard deviation.

The contribution of each connector is made up of the specific contributions of its intermediate processes. Consider, for instance, the environmental impacts associated with the product system of electricity. From the total of 35.49 kg CO_2-eq. (**Mc**) related to the electricity product system, 1.49 kg CO_2-eq. (**Mup**) or 4% (**Rup**) of the impact is caused upstream. The residual difference, 34 kg CO_2-eq., is caused directly during the electricity production process. To put it differently, the elements in **Rup** express the relative loss that would occur if we removed all intermediate processes required by a unit process. For example, deleting all intermediate processes from the production of iron ore (steel and fuel) would cause a decrease of 0%, 85%, 55%, 5%, 3%, and 2% in relation to the total environmental impacts across all product systems in the database.

Note that the connectors perspective results in double counting because the upstream contributions counted in one process (e.g., electricity with 1.49 kg CO_2-eq.) include shares of the contribution already accounted for in the preceding process (e.g., oil with 0.29 kg CO_2-eq.). This is also indicated by the greater-than-one sum of some product systems in **Rup**. Nonetheless, the connector's perspective represents a meaningful viewpoint when focusing on the contributions of the processes throughout all product systems. With this focus, **Rup** allows the identification of consistently important connectors.

2.2.3 Summary measures

Having discussed the calculation procedure for both perspectives, we turn to the question of how the information in the obtained relative contribution matrixes can be investigated.

2.2.3.1 Measures of location and dispersion

The *arithmetic mean* and the *median* are obvious location indicators for the elucidation of the average importance of a process throughout all product systems. We base our analysis mainly on the arithmetic mean because we are interested in typical values representing the center of gravity of the set of contributions associated with a process (Bulmer, 1979, p. 52). Although this gives only a very rough idea of the actual relevance per product system it offers a useful indication of the average importance of a process throughout the database. In addition, we apply the *standard deviation* (SD) to capture the variability among the product specific contributions of a process (Bulmer, 1979, p. 54).

Figure 2-5 (bottom) shows their exemplary application to **Rd**. The mean contribution vector, Ø, shows that electricity has the highest contribution (39%) throughout all product systems followed by oil (33%). This means that if electricity were produced with no the CO_2 and CH_4 emissions, it would decrease the environmental impacts of all product systems by 39%. Producing electricity and oil without CO_2 and CH_4 emissions would, on average remove 72%. The example shows the additive character of contributions viewed from the causer perspective. Consequently, the deletion of all unit processes that incorporate causing elements, i.e., exchanges with the environment, would remove all environmental impacts from the database.

Figure 2-7 (bottom) shows the measures of location for **Rup**. Iron ore and fuel production appear as the largest connectors with a mean contribution of 25% and 24%, respectively. Similar to the causer perspective, this can be interpreted as an average reduction of 25% throughout all product systems if we removed all intermediate processes required by iron ore production. In contrast to the causer perspective, however, cumulating the mean process contributions of the connector perspective results in double counting. This is like measuring the throughput of a water stream before and after it flows together with another stream; the amount of water measured after the junction includes all of the water already accounted for before the junction. Each junction represents a sensitive intersection where one can divert or stop the throughput. Therefore, the cumulative value of 0.49 related to the removal of all intermediate processes required by iron ore and steel must be interpreted as a theoretical maximum that expresses the amount of contribution which, on average, "flows through" each of the connectors in the database when they are measured independently of each other. Consequently, the actual cumulative decrease in environmental impacts of two and more connectors would always be lower than indicated by the cumulative value because interdependencies between the connectors decrease the actual value. Therefore and in contrast to the causer perspective, removing all connectors from the database would not eliminate all environmental impacts.

2.2.3.2 Lorenz curve and related measures of inequality

The *Lorenz curve* is the most popular tool for visualizing and comparing income inequality (Duclos and Araar, 2006, p. 43). It indicates the cumulative percentage of total income held by a cumulative proportion of the population by ranking the income of each individual in increasing order inequality (Duclos and Araar, 2006, p. 43). To the best of our knowledge, the Lorenz curve has so far not been applied in the context of LCA. However as we will see next, it provides some interesting properties for visualizing and analyzing our mean process contributions.

We use the mirror image of the Lorenz curve (MLC) to express the cumulative percentage of the total contribution held by a cumulative proportion of processes. That is, instead of plotting the lowest contribution first, we start with the process whose mean contribution throughout the database is the largest and proceed by adding all other mean process contributions in descending order. More formally, if we have n ordered process contributions, such that $\emptyset_{\mathrm{i}}$ is the mean of process i and $\emptyset_1 \geq \emptyset_2 \geq \cdots \geq \emptyset_{\mathrm{n}}$, then the discrete MLC is defined as the polygon joining the points $\left(\mathrm{p} = \frac{\mathrm{h}}{\mathrm{n}}, \mathrm{c} = \frac{\mathrm{ML_h}}{\mathrm{ML_n}}\right)$ where $\mathrm{h} = 1, 2 \ldots \mathrm{n}$, $\mathrm{p}_0 = 0$, $\mathrm{c}_0 = 0$, $\mathrm{ML_n} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \emptyset_{\mathrm{i}}$ and $\mathrm{ML_h} = \sum_{\mathrm{i}=1}^{\mathrm{h}} \emptyset_{\mathrm{i}}$ (Damgaard and Weiner, 2000).

Figure 2-8 (left) shows the MLCs obtained from the mean process contribution vector of **Rd** and **Rup**, and the diagonal of the unit square reflecting the line of perfect equality; if each process has the same mean process contribution, the MLC would follow this line. The convexity of the MLC reveals the density of the mean process contributions at various percentiles, p. The density is larger for higher values of p since, as p increases, the slope changes less rapidly (Duclos and Araar, 2006, p. 44).

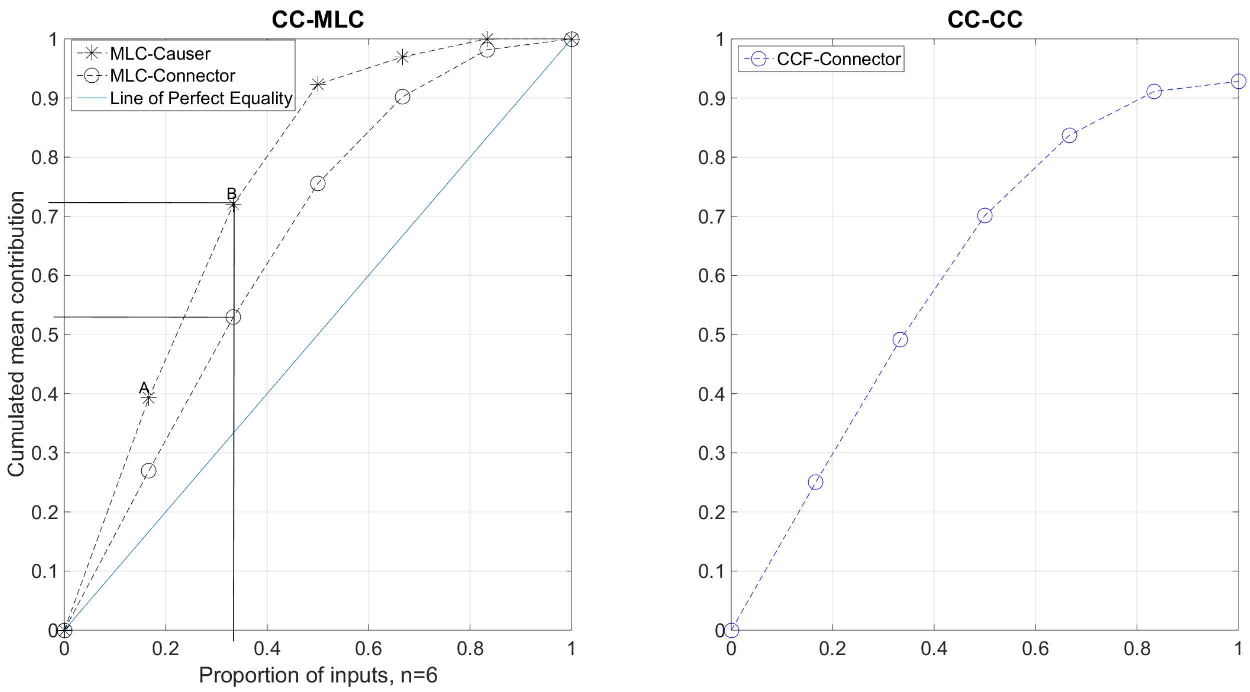


Figure 2-8: Left: Mirrored Lorenz curves (MLC) obtained from the mean process contribution vector; each mean process contribution is normalized to the sum of all mean process contributions. The diagonal of the unit square

represents the line of perfect equality. Right: The cumulated contribution (CC) curve obtained from ML_h computed from the descending sorted mean process contribution vector Ø of Rup.

The Cumulated Contribution Function (CCF) (Figure 2-8, right) shows the cumulated proportion of the mean contribution in $\mathrm{ML_h}$, which is transmitted by a cumulated proportion of the connectors. The only difference to the MLC is that the values are not normalized according to the total contribution of all connectors ($\mathrm{ML_n}$). It shows the cumulated mean "throughput" transmitted by h connectors throughout all product systems. That is, it specifies the potential contribution one can impact by looking at the connectors only.

The total amount of inequality in the MLC can be summarized by the *Gini coefficient* (GC) which is well known for its use for the expression of inequality of income or wealth (Slottje et al., 2001). The GC graphically expresses the ratio between

- the area between the MLC and the diagonal of the unit square (which reflects a perfect equal distribution) and
- the overall area above the diagonal of the unit square, i.e., 0.5.

The more equal the distribution of the relative mean contributions are, the more will the MLC approach the diagonal and consequently the smaller the area and the resulting GC. Equation 11 describes the computation of GC from unordered data as the "relative mean difference, i.e., the mean of the differences between every possible pair of processes, divided by the mean size" (Damgaard and Weiner, 2000, p. 1139).

$$GC = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|\mathrm{Ø_i}-\mathrm{Ø_j}|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Ø_j}} \tag{11}$$

The GC can take values between 0, when all mean contributions in Ø are equal, and 1 when every process except one has a mean contribution of zero (Damgaard and Weiner, 2000). Expressing the inequality in the MLC by means of the GC gives us an indication of the utility of prioritization. A GC of almost zero would indicate that prioritization is meaningless as each unit process is of equal importance. In this case, random sampling and improving unit processes would provide equal utility. Vice versa, a GC of 1 would indicate that prioritization is of the highest possible utility as only one unit process in the database is responsible for all contributions throughout the database. For our CE the GC is 0.50 and 0.31 for **Rd** and **Rup**, respectively, showing that (i) prioritization has merit because there is some inequality in the mean process contribution of the causer perspective and (ii) the mean contributions in the connector perspective are more equally distributed.

The GC does not express all information in the MLC (Damgaard and Weiner, 2000), and we therefore use additional summary measures to highlight further important aspects. The *Concentration Ratio (CR)* measures the share in total contribution held by the largest h processes of the population; the larger that share, the more unequal is the distribution among the mean process contributions. It offers a "snapshot" of the cumulated distribution at rank h (equation 12). For example CR(2) is obtained by summing up the two largest mean process contributions, i.e., electricity and oil for **Rd** (0.72) and iron ore and fuel for **Rup** (0.53)—see the straight lines in Figure 2-8 for a visual representation.

$$\mathrm{CR(h)} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{h}} \emptyset_{\mathrm{i}}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \emptyset_{\mathrm{i}}} \quad \text{with } \emptyset_1 \geq \emptyset_2 \geq \cdots \geq \emptyset_{\mathrm{n}} \tag{12}$$

The inverse to CR(h), i.e., h(t), measures the amount of processes h required to exceed a certain threshold, t of the overall mean process contribution. Although the effort or time required to review a process is subject to large fluctuations, the amount of processes can be considered as a rough proxy for the effort required for reviewing a certain proportion of the MLC. We compute h(t) by determining the cardinality, expressed as n(), of a vector l containing the set of cumulated mean contributions c, that are lower than the threshold t. Since l will always start with zero (recall that $c_0 = 0$), the cardinality of I already indicates the amount of all processes h, whose accumulated contribution is larger than but minimal to the threshold (equation 13).

$$\mathrm{h(t)} = \mathrm{n}(l); \text{with } l_i = c_i \text{ for } c_i < \mathrm{t} \tag{13}$$

For **Rd**, h(50%) would amount to 2 since the accumulated contribution of electricity and oil production is required to exceed the threshold of 50%. For **Rup**, h(50%) would amount to 2 since the accumulated contribution of iron ore and fuel production are required to exceed the threshold.

The MLC portrays the entire distribution of mean-normalized contributions (Duclos and Araar, 2006, p. 44). For example, the slope of 1.96 between A and B in Figure 2-8 indicates that the contribution of the largest process in **Rd** (steel) is 1.96 times higher than μ, the mean of all process contributions. That is, we will find μ at that percentile at which the slope s of MLC(p) equals one (Duclos and Araar, 2006, p. 45). This property allows the identification of an important turning point in the MLC. All mean process contributions before this point are higher than μ; all following contributions are lower. We use this property to identify $h_{>\mu}$, the amount of the processes in the population that have a larger mean contribution than μ throughout the database, respectively (equation 14). We first compute the slope s for each segment of the MLC and then determine the cardinality of a vector l that contains all segments whose slope is greater than one. For the mean process contribution of **Rd**, $h_{>\mu}$ would be approximated with 3 because three out of the 6 mean process contributions in the population have a slope greater than one.

$$h_{>\mu} = n(l); \text{with } l_i = s_i \text{ for } s_i > 1; s_i = \left(\frac{c_{i+1} - c_i}{\frac{1}{n}}\right) \tag{14}$$

Finally, the applicability of these advanced summary measures is not limited to the mean process contributions. As we will show later, they are also useful for summarizing the detailed contribution pattern of each individual process throughout the database (see Table S2-4and Table S2-5 in section 2.9.2.2).

2.3 Application to version 3.1 of the ecoinvent database

2.3.1 Implementation

We implemented the method to analyze version 3.1 of the ecoinvent database in the system model "Allocation, cut-off by classification"[37]. The database covers 11,304 unit processes and represents one of the largest LCI databases. Because our focus is on the proof of concept and not on the result per se, the demonstration presents only the results for two randomly selected LCIA indicators (Table 2-1).

Table 2-1: LCIA indicators used for the analysis of ecoinvent V.3.1 and their abbreviation (Abbr.)

No.	LCIA method	Indicator	Unit	Abbr.
1	CML 2001	Resources, depletion of abiotic resources	kg antimony-eq.	ARD
2	ReCiPe Midpoint (H)	Photochemical oxidant formation	kg NMVOC	POFP

The detailed computation procedure can be described as follows: First, we compute **Rd** and **Rup** (equation 1-10). That is, for each LCIA indicator we generate two matrixes with a size of 11,304 x 11,304 or roughly 256 million potential data points. In order to give an impression of this data, Figure Figure S2-14 (in section 2.9) exemplarily illustrates **Rd** and **Rup** for ARD.

Second, we calculate the location and dispersion measures for each process listed in **Rd** and **Rup**. In order to facilitate a richer interpretation, we construct an MLC from the individual contribution patterns of each process and apply the elaborated measures of inequality (equation11-14). We also calculate the quantity of contributing processes (QCI) in the mean contribution vectors by replacing all non-zero elements with one. The QCI is the sum of this vector. It tells us the actual quantity of processes that include at least one element (elementary flow or intermediate process) which cause or connect to an environmental impact. Finally, we use the mean process contribution vectors, as the basis for the construction of an MLC and then apply the inequality measures again (equation 11-14).

[37] The system model is based on the Cut-off approach where primary (first) production of materials is always allocated to the primary user of a material. Furthermore, a primary producer of a recyclable material does not receive any credit for its provision. Therefore, recyclable materials are available burden-free to recycling processes, and secondary (recycled) materials bear only the impacts of the recycling processes (Ecoinvent, 2015).

2.3.2 Results

Figure 2-9 shows the MLC and the CCF for all of the LCIA indicators assessed. To improve readability we used a logarithmic scale for the share of processes. Table S2-4 and Table S2-5 in the SI show the detailed use and contributions patterns of the ten largest processes for both perspectives and both LCIA indicators.

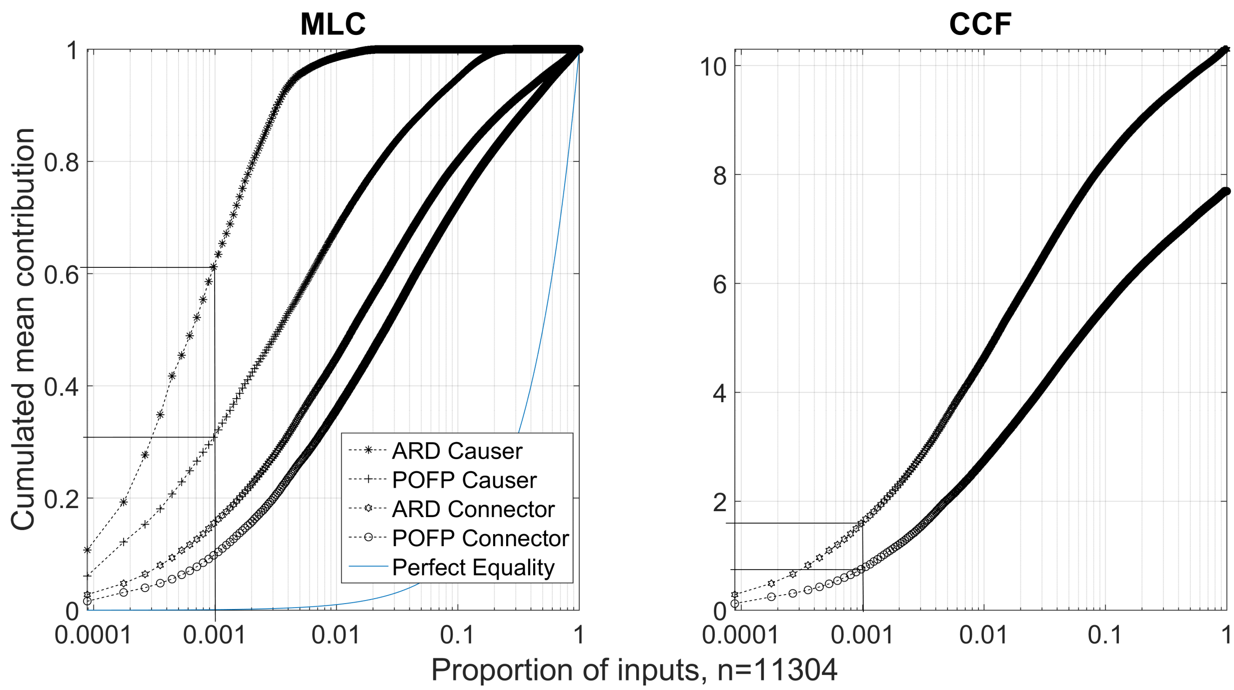


Figure 2-9: MLC and CCF for the LCIA indicators assessed. The CCF curves are obtained from the accumulation of the ordered, mean process contribution of **Rup**.

The inequalities in the mean process contributions are remarkable; 1‰ (or CR(11)) of the largest causers already cumulate between 31% (POFP) and 62% (ARD) of the overall contribution. For both LCIA indicators, the connectors' MLCs show less concentration than the causers' MLCs; the 11 largest connectors transmit 10% (POFP) and 15% (ARD) of the contribution transmitted via all connectors, respectively. However, in terms of actual contribution (see CCF), this equals 75% (POFP) and 159% (ARD). That is, on average, the major connectors relocate more contribution than caused by the major causers. The mean overall throughput transmitted by all connectors amounts to 10.3 (ARD) and 7.7 (POFP) times the total contribution.

Table 2-2 shows further summary measures for the presented MLC.

Table 2-2: Summary statistics for the LCIA indicators and perspectives applied.CR: concentration ratio; GC: gini coefficient; h: amount of processes exceeding a certain threshold; QCI: quantity of contributing processes in the mean contribution vector.

Perspectives of analysis	**Causer**		**Connector**	
LCIA indicator	**ARD**	**POFP**	**ARD**	**POFP**

CR(7)	0.49	0.25	0.12	0.07
GC	0.9973	0.9634	0.8307	0.7715
$h_{>\mu}$ [$CR_{(h>\mu)}$]	192 [0.99]	976 [0.94]	1,302 [0.82]	1,576 [0.78]
h(60%)	11	70	288	509
h(80%)	24	261	1,126	1,895
QCI	368	3,369	10,859	10,873

The GC confirms the high inequality in the size classes of the mean process contributions and indicates a high utility for prioritization across both LCIA indicators and perspectives. The rank of the processes in the population that have a larger contribution than μ, $h_{>\mu}$, and the corresponding $CR(h_{>\mu})$ helps us in the interpretation of the GC. The higher the amount of processes $h_{>\mu}$ and the lower the corresponding $CR(h_{>\mu})$, the less is the relative contribution of an process concentrated and consequently the lower the GC.

$h_{60\%}$ and $h_{80\%}$ reveal how many processes we have to review to exceed a certain proportion of the overall contribution. The amount of processes required differs substantially across LCIA indicators but is mainly dependent on the perspectives of analysis. For example, to exceed a contribution of 80% we have to focus on 24 processes for ARD. Yet, to exceed the same contribution in the connector perspective we have to review 1,126 processes. The QCI tells us why: almost all processes in the database include connecting elements, i.e., link to an intermediate process, but only some processes actually include causing elements, i.e., an elementary flow that is relevant for a particular LCIA method. For example, only 368 processes include an elementary flow relevant for the LCIA indicator ARD but 10,859 processes link to an intermediate process with an environmental impact generated by a causer upstream. Even though the inequality in the mean connector contributions is still remarkably high, e.g., 3%-5% of the connectors are required to cumulate a contribution of 60%, the amount of processes required to review a certain proportion of their overall contribution is much higher than in the causer perspective.

2.4 Evaluation

The goal of this section is to evaluate our method with regard to its validity, performance, and utility. Our evaluation is based on the experience gained in the demonstrations.

2.4.1 Validity

Validity means that “the method works and does what it is meant to do” (Gregor and Hevner, 2013, p. 351). The successful application to the ecoinvent database shows that our method is operational and enables the identification of the unit processes to be prioritized for improvement. It also gives an overall indication of the effort (amount of processes) required to review a certain proportion of the overall contribution (benefit). That is, the method achieved its purpose in the application domain of relevance.

2.4.2 Performance

The sequence of our computation algorithm requires just one matrix inversion, i.e. the calculation of **S**, during the calculation for the first LCIA indicator. All subsequent LCIA indicators are computed on the basis of **S**. This allows efficient computation. A modern[38] laptop computer requires roughly 2 minutes for the calculation of all summary measures associated with one LCIA indicator.

2.4.3 Utility: Causer and connector perspective

Figure 2-10 shows that there is a correlation between the causer and the connector perspective, i.e., many product systems consistently call unit processes which include "connecting" and "causing" elements of relevance and therefore one and the same unit process can be important in both perspectives (B). At the same time, the most important connector, the "global market for petroleum" (circled) has no contribution in the causer perspective (A)—and therefore does not appear in the logarithmic illustration. This holds true for many other connectors as well.

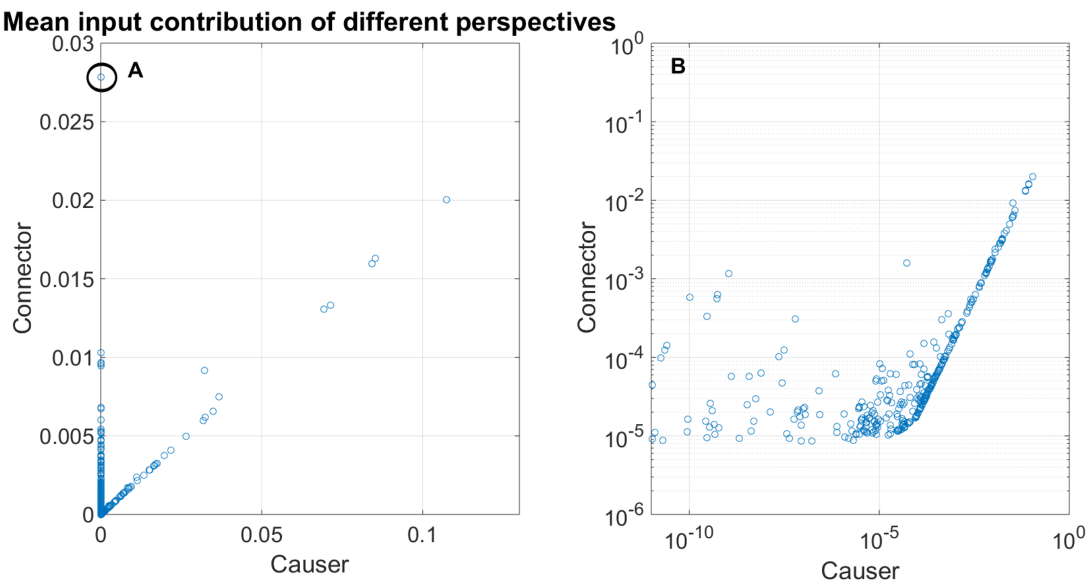


Figure 2-10: Scatter plot illustrating the mean process contribution of the ARD causer vs. the ARD connector perspective with linear (A) and logarithmic (B) scales.

The connector and causer perspectives reveal two different capacities of one and the same unit process. The distinction is useful because it facilitates a more fine-grained configuration of improvements. For example, the finding that the major connectors often "control" more relative contribution than the major causers suggests that they should receive more attention.

[38] i7-4800MQ CPU @ 2.7 GHz, 16 GB RAM

2.4.4 Utility: Prioritization in the application domain

Our application domain is the organizational and technical system that develops and maintains LCI databases. In order to assess the general utility of prioritizations in this domain, we compare the performance of our approach with the common practice. To our knowledge, none of the LCI databases applies operational methods to prioritize LCI database improvements. In light of the lack of any operational method, it seems reasonable to assume uniform sampling as a simple, first estimation. This means that each unit process has the same probability to be selected for improvement. Figure 2-11 shows the cumulated mean contribution reviewed (benefit) as a function of the required proportion of processes (effort) for the LCIA indicator POFP causer; once for our prioritized approach (MLC) and once for uniform sampling.

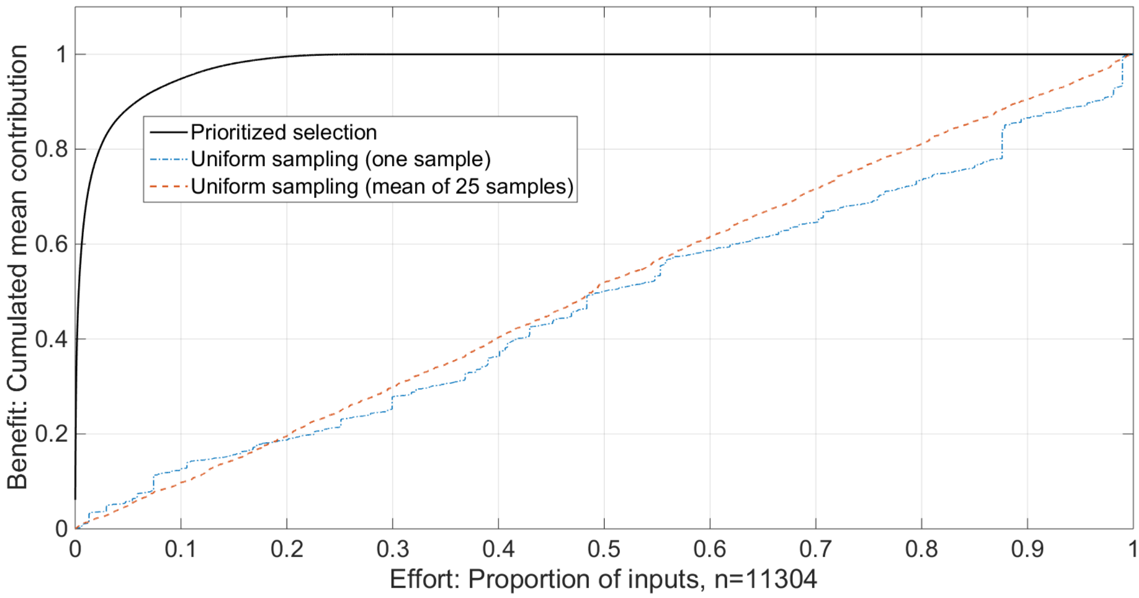


Figure 2-11: Comparison of benefit (cumulated mean contribution reviewed) of a prioritized mean contribution vector and a uniformly sampled mean contribution vector as a function of the effort (the proportion of processes reviewed).

As shown in Figure 2-11, the performance of a uniform sampling approach approximates the line of perfect equality. Even though the assumption of uniform sampling is simplistic (and expert-based prioritization might perform better), the comparison allows some general reasoning about and gives a first indication of the utility of prioritization in the application domain.

- First, the higher the inequality in the mean process contribution, the more will the MLC depart from the line of perfect equality (uniform sampling) and consequently the higher the utility of prioritization. That is, our measure of the utility of prioritization, the Gini coefficient, implicitly assumes uniform sampling as the baseline or reference system.
- Second, the utility of prioritization is highest for the largest process and decreases rapidly. For example, uniform sampling would cause 664 times more effort to review the contribution held by

the largest process of POFP (6%), and 219 times more effort to review the contribution held by the twenty largest processes of POFP.

Bearing this in mind, the MLC can be considered as a powerful tool for the analysis of the inequality in the mean process contributions.

2.4.5 Utility: Prioritization across many LCIA indicators

Figure 2-12 contrasts the mean causer contribution according to ARD and POFP in a linear (A) and logarithmic fashion (B). It underlines two important insights. First, there is moderate correlation within the prioritization space of both LCIA indicators, i.e., different LCIA indicators will point to the same unit processes (B).

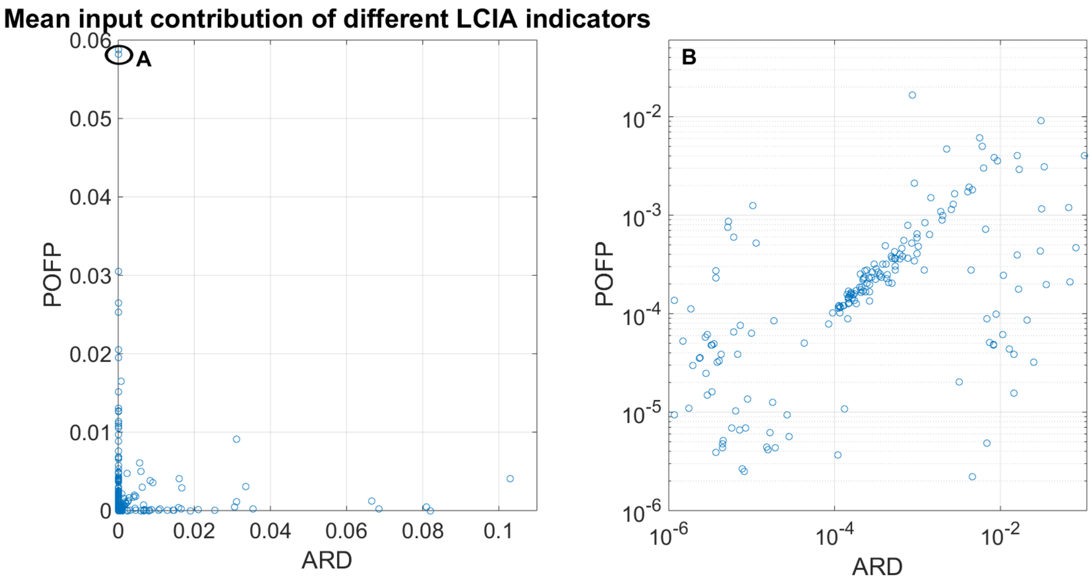


Figure 2-12: Scatter plot illustrating the mean process contribution of ARD vs. POFP with linear (A) and logarithmic (B) scales.

At the same time, different LCIA indicators have different inventory support and thus prioritize different unit processes (A). For example, the second largest process of POFP, “diesel, burned in building machine (GLO)” (circled), has zero relevance for ARD.

This finding indicates that “robust prioritization” requires computation across many LCIA indicators, simply because every additional indicator adds a new set of unit processes into the prioritization space. We already showed that prioritization has merit for one selected LCIA indicator. Yet the ecoinvent database offers hundreds of LCIA indicators. It is therefore a valid question whether prioritizing across many LCIA indicators still provides any benefit. In order to address this question we prioritized across all modern LCIA methods provided by the ecoinvent database, i.e., 8 LCIA methods in the form of 112 LCIA indicators (see Table S2-6 in section 2.9.3 for the detailed selection). Figure 2-13 shows the effective amount of processes that have to be reviewed as a function of the LCIA indicators considered and for different thresholds. The SI (Figure S2-15) presents the algorithm in detail.

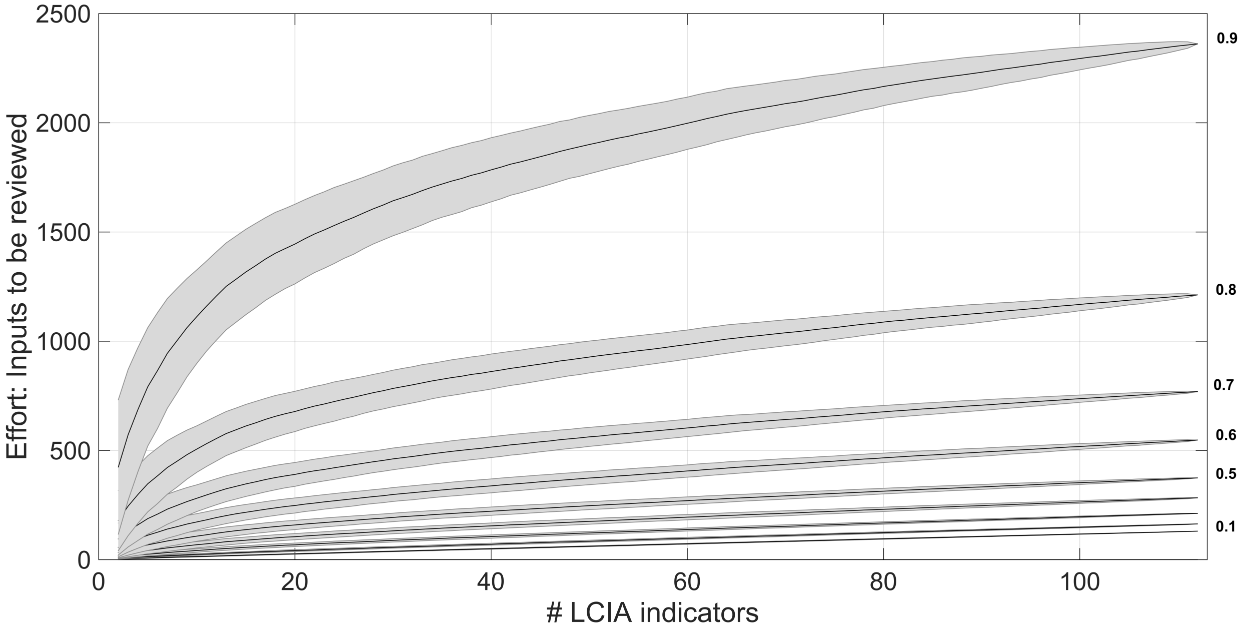


Figure 2-13: Processes to be reviewed as a function of the LCIA indicators considered for different thresholds. Reading example. To review 80% of the contribution across all 112 LCIA indicators we require on average 1,221 (out of 11,304) processes. It is the sampled sequence of the LCIA indicator which causes the variation. The black lines represent the arithmetic mean of 1,000 uniform samples. The gray lines indicate the corresponding standard deviation. The SI (Figure S2-16) shows the corresponding illustration for the connector perspective.

Figure 2-13 tells us three notable things. First, even when applied across 112 LCIA indicators, prioritization has considerable utility; roughly 5% of the unit processes in the database are sufficient to review 60% of the overall contribution across all LCIA indicators. The reason is the high correlation among LCIA indicators, which results in a large amount of overlap in the prioritization space. Second, the amount of "unknown" processes inserted into the prioritization space with each additional LCIA indicator flattens out. Consequently, using more than approx. 60 LCIA indicators will not add many more processes into the prioritization space. Finally, as indicated by the decrease in the standard deviation, the sequence of LCIA indicator selection matters, particularly at the beginning.

It is noteworthy, however, that this approach indicates something like an upper limit of the number of processes which need to be considered because it ignores the scientific quality of and stakeholder interest in LCIA indicators. Filtering available LCIA indicators according to both aspects will reduce the actual number of processes in the prioritization space.

2.5 Discussion

Precise knowledge of the elements with the largest contribution is a prerequisite for the systematic improvement of LCI databases. Our method contributes to this end with two novel solutions; (i) a mathematical, computational framework that allows the application of CA to entire LCI databases for two different perspectives and (ii) summary statistics that facilitate the aggregation and interpretation of the data. The contribution-based approach is applicable to any disaggregated LCI database that provides, or can be transformed into, the technosphere, biosphere, and characterization matrix.

At a higher level of abstraction, this approach could be transferred to any problem of system modeling where models of interacting parts of the system are given and the question arises, which of these models should be improved with the highest priority with regard to a set of indicators computed with the overall model. This type of problem can occur, e.g., in large system dynamics models that are reusing stereotyped submodels (Ahmadi Achachlouei and Hilty, 2015), in agent-based modeling (Kostadinov et al., 2014), or in material flow modelling (Bornhöft et al., 2016).

Returning to the domain of LCI databases, some important limitations of our approach need to be considered. First, each product system is given equal weight, meaning that we assume a uniform distribution of importance across all products in in the database. Similar products (such as the numerous electricity markets) use similar product systems and therefore augment the relative importance of the same processes. Consequently, our assessment of the most important processes is determined by the available product systems and their structure as it has been modelled. This is intended and useful for the contribution-based elucidation of the inherent structural dependencies in the existing network structure. Yet this inward perspective provides no direct support for the identification of "blank spots" and is therefore unable to direct research efforts to economic sectors that may be underrepresented. In fact, it focuses exclusively on the assessment of the status quo. This shows the iterative character of the method: with each new version of a database, our method should be applied anew.

Second, our study employs simple summary indicators to reflect complex issues. Equating "effort" with the number of unit processes assumes uniform data collection challenges for all processes. Even though this ignores the large heterogeneity among unit processes, we expect, as Majeau-Bettez (2011) does, that "the great number of processes in the ecoinvent database allowed for an averaging out of extremes and a reliable 'average inventory effort' indicator" (Majeau-Bettez et al., 2011, p. 10175). It is also important to keep in mind that our measure of benefit, the cumulated contribution reviewed, does not quantify the actual improvement potential. Identifying the actual improvement potential requires matching the contribution dimension with the uncertainty dimension or, more specifically, the quantification of the mean contribution to uncertainty of a process throughout all product systems (Heijungs and Kleijn, 2001). The integration of the uncertainty dimension would support a much more fine-grained configuration of improvements and therefore further advance the "degree to which the method achieves its desired benefit in practice," i.e., its effectiveness (Venable et al., 2012, p. 426).

Third, there are several elements of interest along which to decompose the results. Our elements of interest were the unit processes with regard to their relative contribution to a particular LCIA indicator. Such a CA reveals the relative relevance of each process throughout the overall database. However, further aggregation (into sectors of geographies) or disaggregation (into the underlying elementary and intermediate flows) is possible and worthwhile, although outside the scope of this paper.

Finally, the MLCs are based on the arithmetic mean as its property to represent the "real" balance point of a population allows the meaningful quantification of the average contribution transmitted via the connectors (the CCF). The MLC analysis and related statistics can easily be switched to the median for the purpose of sensitivity analysis. Generally, the MLCs based on the median show a larger degree of inequality because important processes remain important and unimportant processes tend to become even more unimportant. The data outputs of the instantiation (see Table S2-4 and Table S2-5 in section 2.9.2.2) allow dynamic analysis of each process using data mining tools such as pivot tables.

We were not able to identify similar formal approaches to focusing improvement efforts in LCI databases in the literature. Some authors, however, address a similar end using different means (Majeau-Bettez et al., 2011) or a different end using similar means (Frischknecht et al., 2007). We will briefly describe these approaches.

The prioritization method presented by Majeau-Bettez et al. (2011) identifies the sectors of the economy that are most underrepresented in the version 2.1 of the ecoinvent database. This is an external prioritization; the optimal allocation of the available research capital follows the economic and environmental relevance obtained from the EEIO database. Our approach, in turn, represents an internal optimization where the optimal allocation follows the inherent structural importance of the unit processes according to different LCIA methods. Note that this also offers an improved basis for external optimization because it allows the identification of structural discrepancies between the LCI database and the EEIO. The two approaches represent complementary lenses towards the same end, namely more effective allocation of the available research capital. In this regard, the approach developed by Majea-Bettez et al. (2011) provides guidance for the new accumulation of unit process data, whereas our approach emphasizes the data elements in the present data structure which should receive the most attention.

We developed our method for the prioritization of improvement efforts in LCI databases, but the contribution-based approach also has value outside of this context. First, it can be used to improve performance in dealing with all issues concerned with the relative environmental relevance of particular processes, economic sectors, or product categories. For example, Frischknecht et al. (2007) analyzed the relative environmental relevance of capital goods throughout the ecoinvent database to decide whether capital goods can be excluded. Such questions can be addressed very quickly on the basis of the relative contribution matrixes. To give an idea of the possibilities, we show in the SI an analysis of the most important sectors (see Figure S2-17 in section 2.9.4.1) as an example. Second, it offers interesting starting points for systematic plausibility checks and error analyses, not only with regard to LCI databases but also for LCIA indicators. For example, Rørbech et al. (2014) evaluate the relative importance of elementary flows across different resource depletion LCIA indicators. In this domain, our method offers important and continuous feedback for LCIA method developers because it allows the detailed assessment of the inventory support for different LCIA methods. Third, the method provides valuable information for the generation of new LCI databases because it facilitates the systematic and robust identification of the "core" of large LCI databases. For example, (2013) used a contribution-based approach to identify potential datasets for recontextualization (see section 2.1.4.1). Our formalized approach facilitates a more fine-grained analysis for this type of investigation. Finally, the relative contribution matrixes might provide a consistent basis for topological network analysis. Although the generation of such weighted networks is considered straightforward for the analysis of dependencies among interacting parts of a system, others have "not yet been able to identify a logical and consistent way to do so" (Heijungs, 2015, p. 162). The relative contribution in our matrixes could represent the missing "common currency" required for this task.

2.6 Conclusion

We presented an operational contribution-based approach that assists organizations maintaining LCI databases in the identification of the elements to be prioritized for improvement. Our approach supports the systematic and fine-grained configuration of improvements from an inward perspective. At a more abstract level, our approach can be used to allocate research resources among large sets of interacting elements or submodels that are used as building blocks in higher-level system models.

The demonstration and evaluation using the case of the ecoinvent database shows that the contribution-based approach has merit; we provide evidence for the fact that a tiny and robust set of unit processes is responsible for the lion's share of contribution across the entire LCI database and even across many LCIA indicators. This suggests that the effective and efficient improvement of the entire LCI database is feasible and worthwhile. The method developed will be applied in the development of priorities for the updating and maintenance of the ecoinvent database as one contributor to the list of priorities in database development.

10,000 datasets sounds like a large number, but it is no match for the complex economic system LCA strives to model from the bottom up. To succeed, the LCA field requires a more deliberate strategy for the gathering, structuring, and accumulation of LCI data.

2.7 Acknowledgements

The authors express their gratitude to Hans-Jörg Althaus, who provided early comments and Nikolaus Bornhöft and Andi H. Widok for prove reading.

2.8 References

Ahmadi Achachlouei, M., Hilty, L.M., 2015. Modeling the Effects of ICT on Environmental Sustainability: Revisiting a System Dynamics Model Developed for the European Commission, in: Hilty, M.L., Aebischer, B. (Eds.), ICT Innovations for Sustainability. Springer International Publishing, Cham, pp. 449–474.

Bornhöft, N.A., Sun, T.Y., Hilty, L.M., Nowack, B., 2016. A dynamic probabilistic material flow modeling method. Environ. Model. Softw. 76, 69–80. doi:10.1016/j.envsoft.2015.11.012

Bourgault, G., Lesage, P., Samson, R., 2012. Systematic disaggregation: a hybrid LCI computation algorithm enhancing interpretation phase in LCA. Int. J. Life Cycle Assess. 17, 774–786. doi:10.1007/s11367-012-0418-7

Bulmer, M.G., 1979. Principles of Statistics. Dover Publications, Inc., New York.

Ciroth, A., 2013. OpenLCA - Process contribution - Documentation [WWW Document]. URL http://openlca.org/documentation/index.php/Process_contributions (accessed 8.16.13).

Ciroth, A., Fleischer, G., Steinbach, J., 2004. Uncertainty calculation in life cycle assessments. Int. J. Life Cycle Assess. 9, 216–226. doi:10.1007/BF02978597

Damgaard, C., Weiner, J., 2000. DESCRIBING INEQUALITY IN PLANT SIZE OR FECUNDITY. Ecology 81, 1139–1142. doi:10.1890/0012-9658(2000)081[1139:DIIPSO]2.0.CO;2

Duclos, J.-Y., Araar, A., 2006. Poverty and Equity: Measurement, Policy, and Estimation with DAD. Springer.

Ebner, F., 2013. Results, Interpretation and Reporting: GaBi Software [WWW Document]. Gabi-Softw. URL http://www.gabi-software.com/software/gabi-software/gabi-5/functionalities/results-interpretation-and-reporting/ (accessed 8.16.13).

Ecoinvent, 2015. Allocation cut-off by classification [WWW Document]. URL http://www.ecoinvent.org/database/ecoinvent-version-3/system-models/allocation-cut-off-by-classification/ (accessed 4.16.15).

Finnveden, G., Hauschild, M.Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., Suh, S., 2009. Recent developments in Life Cycle Assessment. J. Environ. Manage. 91, 1–21. doi:10.1016/j.jenvman.2009.06.018

Frischknecht, R., Althaus, H.-J., Bauer, C., Doka, G., Heck, T., Jungbluth, N., Kellenberger, D., Nemecek, T., 2007. The environmental relevance of capital goods in life cycle assessments of products and services. Int. J. Life Cycle Assess. 12, 7–17.

Goedkoop, M., Oele, M., 2004. SimaPro 6 - Introduction to LCA with SimaPro.

Gregor, S., Hevner, A.R., 2013. Positioning and Presenting Design Science Research for Maximum Impact. MIS Q. 37, 337–355.

Groen, E.A., Heijungs, R., Bokkers, E.A.M., de Boer, I.J.M., 2014. Methods for uncertainty propagation in life cycle assessment. Environ. Model. Softw. 62, 316–325. doi:10.1016/j.envsoft.2014.10.006

Guinée, J.B., Gorrée, M., Heijungs, R., Huppes, G., Kleijn, R., de Koning, A., van Oers, L., Sleeswijk, A.W., Suh, S., de Haes, H.A.U., de Bruijn, H., van Duin, R., Huijbregts, M.A.J., 2001. Life Cycle Assessment. An operational guide to the ISO standards.

Heijungs, R., 2015. Topological Network Theory and Its Application to LCA and IOA and Related Industrial Ecology Tools: Principles and Promise. J. Environ. Acccounting Manag. Volume 3, pp. 151–167. doi:10.5890/JEAM.2015.06.005

Heijungs, R., 2012. Spatial differentiation, GIS-based regionalization, hyperregionalization, and the boundaries of LCA., in: Environment and Energy (Editorial Series of Italian Commodity Science Academy and Engineering Association of Messina). FrancoAngeli, Milano, Italy, pp. 165–176.

Heijungs, R., 1996. Identification of key issues for further investigation in improving the reliability of life-cycle assessments. J. Clean. Prod. 4, 159–166. doi:10.1016/S0959-6526(96)00042-X

Heijungs, R., Huijbregts, M., 2004. A review of approaches to treat uncertainty in LCA. Complex Integr Resour Manag Proc 2nd Bienn Meet Int Env. Model Softw Soc.

Heijungs, R., Kleijn, R., 2001. Numerical approaches towards life cycle interpretation five examples. Int. J. Life Cycle Assess. 6, 141–148. doi:10.1007/BF02978732

Heijungs, R., Suh, S., 2002. The computational structure of life cycle assessment. Kluwer Academic Publishers, Dordrecht.

Heijungs, R., Suh, S., Kleijn, R., 2005. Numerical Approaches to Life Cycle Interpretation - The case of the Ecoinvent'96 database (10 pp). Int. J. Life Cycle Assess. 10, 103–112. doi:10.1065/lca2004.06.161

Hellweg, S., Milà i Canals, L., 2014. Emerging approaches, challenges and opportunities in life cycle assessment. Science 344, 1109–1113.

Huijbregts, M.J., 1998. Application of uncertainty and variability in LCA. Int. J. Life Cycle Assess. 3, 273–280. doi:10.1007/BF02979835

ISO, 2006. ISO 14044. Environmental management - Life cylce assessment - Requirements and guidelines.

Kostadinov, F., Holm, S., Steubing, B., Thees, O., Lemm, R., 2014. Simulation of a Swiss wood fuel and roundwood market: An explorative study in agent-based modeling. For. Policy Econ. 38, 105–118. doi:10.1016/j.forpol.2013.08.001

Lesage, P., Samson, R., 2013. The Quebec Life Cycle Inventory Database Project. Int. J. Life Cycle Assess. 1–8. doi:10.1007/s11367-013-0593-1

Majeau-Bettez, G., Strømman, A.H., Hertwich, E.G., 2011. Evaluation of Process- and Input–Output-based Life Cycle Inventory Data with Regard to Truncation and Aggregation Issues. Environ. Sci. Technol. 45, 10170–10177. doi:10.1021/es201308x

Mutel, C.L., Hellweg, S., 2009. Regionalized Life Cycle Assessment: Computational Methodology and Application to Inventory Databases. Environ. Sci. Technol. 43, 5797–5803. doi:10.1021/es803002j

Rørbech, J.T., Vadenbo, C., Hellweg, S., Astrup, T.F., 2014. Impact Assessment of Abiotic Resources in LCA: Quantitative Comparison of Selected Characterization Models. Environ. Sci. Technol. 48, 11072–11081. doi:10.1021/es5023976

Slottje, D., Nieswiadomy, M., Redfearn, M., 2001. Economic inequality and the environment. Environ. Model. Socioecon. 16, 183–194. doi:10.1016/S1364-8152(00)00081-5

Sonnemann, G., Vigon, B. (eds)., 2011. Global Guidance Principles for Life Cycle Assessment Databases: A Basis for Greener Processes and Products (Shonan Guidance Principles).

Tillman, A.-M., 2000. Significance of decision-making for LCA methodology. Environ. Impact Assess. Rev. 20, 113–123. doi:10.1016/S0195-9255(99)00035-9

Venable, J., Pries-Heje, J., Baskerville, R., 2012. A Comprehensive Framework for Evaluation in Design Science Research, in: Peffers, K., Rothenberger, M., Kuechler, B. (Eds.), Design Science Research in Information Systems. Advances in Theory and Practice, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 423–438.

Weidema, B., Bauer, C., Hischier, R., Mutel, C., Nemecek, T., Reinhard, J., Vadenbo, C., Wernet, G., 2013. Overview and methodology. Data quality guideline for the ecoinvent database version 3.

Wernet, G., 2012. ecoinvent v3 - Internationalization of the data supply and international collaborations. Presentation at DF48, Dübendorf.

2.9 Supplementary information I

2.9.1 Materials and method

2.9.1.1 Table of matrixes and vectors

Table S2-3: Matrixes and vectors required to calculate relative contribution matrixes.

No.	Name	Symbol	Content	Dimension	Source/Computation
1	Technosphere matrix	$\mathbf{A}$	Intermediate product consumptions by processes.	$n \times n$	Supplied by the database
2	Biosphere matrix	$\mathbf{B}$	Elementary exchanges (emissions and natural resources) by processes.	$i \times n$	Supplied by the database
3	Identity matrix	$\mathbf{I}$	Ones on the main diagonal and zeros elsewhere. Required to establish (explicitly) a unitary demand for each product.	$n \times n$	Set up by LCA practitioner
4	Characterization matrix	$\mathbf{W}$	Characterization factors by elementary exchanges.	$k \times i$	Supplied by the database
5	Supply (or scaling) matrix	$\mathbf{S}$	Process amount (required to satisfy a unitary product demand).by products.	$n \times n$	$\mathbf{S} = \mathbf{A}^{-1}\mathbf{I}$
6	Characterization vector	w	Characterization vector from the characterization matrix $\mathbf{W}$, i.e. one (selected) LCIA indicator.	$1 \times i$	$w = \boldsymbol{W}_i$
7	Direct contribution vector	e	Direct environmental impacts by processes. Represents for each process the environmental impacts caused directly by its elementary flows.	$1 \times n$	$e = w\mathbf{B}$
8	Direct contribution matrix	$\mathbf{Md}$	Direct, process-specific environmental impacts by products. Shows the direct impact of each process throughout all product systems.	$n \times n$	$\mathbf{Md} = \hat{e}\mathbf{S}$
9	Total impact vector	td	Total environmental impact by products. Shows the total impacts over the entire product system, associated with a unitary product demand.	$1 \times n$	$\mathrm{td}_j = \sum_i \mathbf{Md}_{ij}$
10	Expansion vector	$\vec{1}$	A vector of ones used to expand vectors td and p to matrix form.	$n \times 1$	Set up by LCA practitioner
11	Relative direct contribution matrix	$\mathbf{Rd}$	Relative, process-specific contributions by products. Shows the relative contribution of each process throughout all product systems.	$n \times n$	$\mathbf{Rd} = \mathbf{Md} \oslash (\vec{1}td)$
12	Cumulated impact per process	tn	Cumulated (direct and upstream) environmental impact by processes. Shows the impacts associated with the provision of one unit of each process.	$n \times 1$	$tn = td^{\mathrm{T}} \oslash \hat{\mathbf{S}}$
13	Cumulated contribution matrix	$\mathbf{Mc}$	Cumulated (direct and upstream) environmental impact of processes by products.	$n \times n$	$\mathbf{Mc} = \widehat{tn}\mathbf{S}$
14	Upstream proportion vector	p	Proportion of process-specific environmental impacts caused upstream.	$1 \times n$	$p = 1 - (e \oslash td)$

15	Upstream contribution matrix	$\mathbf{Mup}$	Upstream environmental impact of processes by products. Shows the upstream contribution of each process throughout all product systems.	$n \times n$	$\mathbf{Mup} = \mathbf{Mc} \circ \left(\vec{1}p\right)^{\mathrm{T}}$
16	Relative upstream contribution matrix	$\mathbf{Rup}$	Relative, process-specific upstream contribution by products. Shows the relative upstream contribution of each process throughout all product systems.	$n \times n$	$\mathbf{Rup} = \mathbf{Mup} \oslash \left(\vec{1}td\right)$

2.9.2 Application to version 3.1 of the ecoinvent database

2.9.2.1 The relative contribution matrixes for ARD

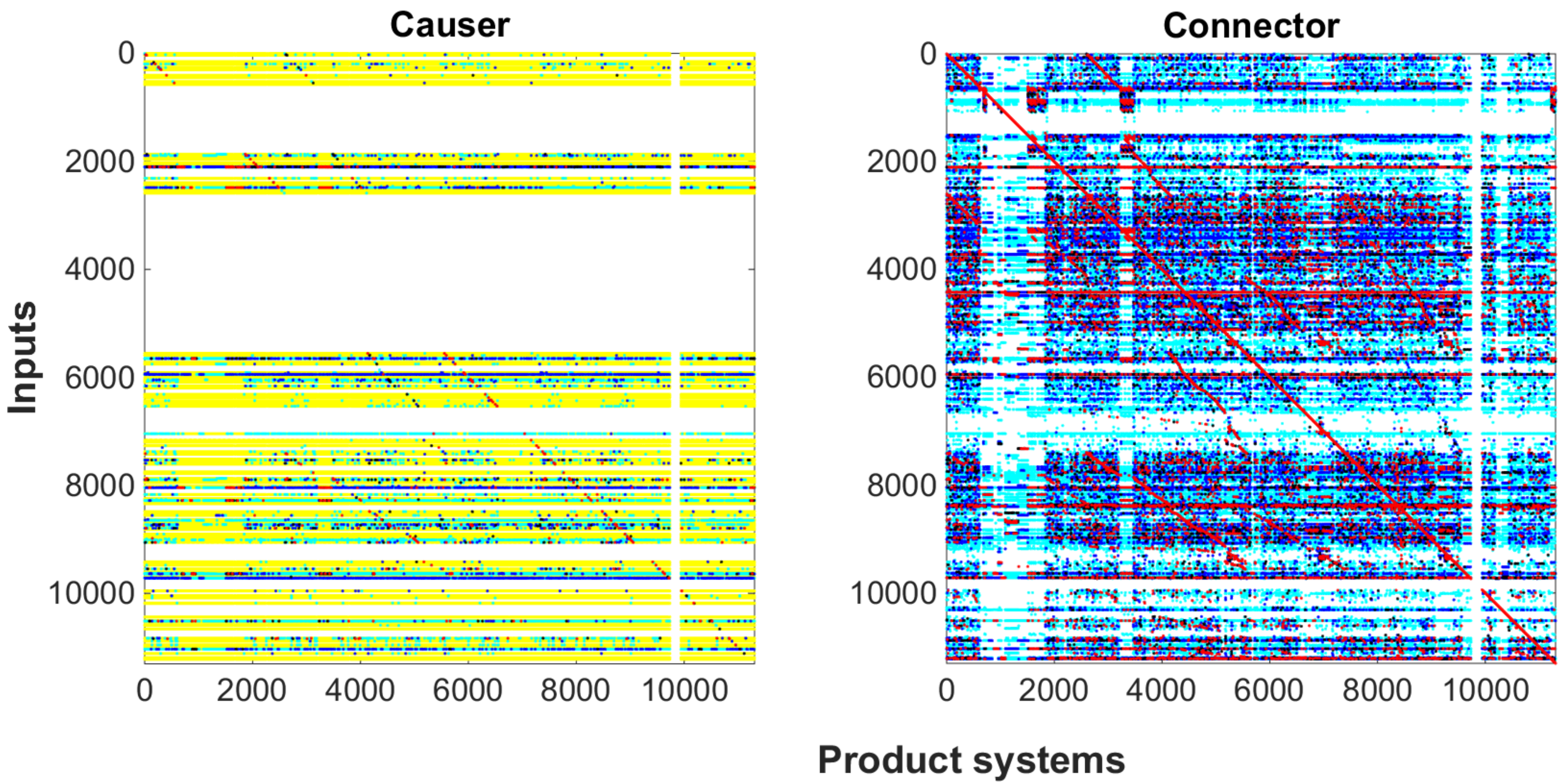


Figure S2-14: Colored-coded spy plots visualizing the use and contribution patterns of 11,304 inputs throughout 11,304 product systems for the relative contribution matrixes Rd and Rup for the LCIA indicator ARD. Color code: contributions exceeding 0%, 1%, 10%, 30%, 50% are yellow, turquoise, blue, black and red, respectively. Inputs with no contribution are shown as white.

2.9.2.2 Use and contribution patterns of the most important inputs

Table S2-4and Table S2-5 show the use and contribution patterns for the ten largest unit processes; once for the causer and once for the connector perspective. The frequency of use (FoU) shows that the ten largest unit processes are used by practically the entire database, i.e., in 96% (~10,850) of the available product systems. To put it differently, an improvement of the largest inputs will affect practically the entire database.

The mean and median of a unit process reveal its average importance across all product systems. A large discrepancy of the median in rank indicates that the mean is affected by some particularly large (or small) contributions, i.e., outliers. CR(80) gives us an indication about the 80 largest contributions. The higher CR(80), the more concentrated are the contributions of a unit process

The Gini coefficient (GC) expresses the inequality in the contributions throughout all product systems; extremely heterogeneous contributions are indicated by a high GC, very homogenous contributions by a low GC. The $h_{>\mu}$ and the corresponding CR($h_{>\mu}$) provides further information for the interpretation of the GC. The higher the amount of inputs $h_{>\mu}$ and the lower the corresponding CR($h_{>\mu}$), the less is the relative contribution of a unit process concentrated and consequently the lower the GC. Note that the standard deviation is not very sensitive to these kind of variations.

Table S2-4: Use and contribution patterns of the most important unit processes according to the causer perspective for ARD and POFPs. FoU, frequency of use; SD, standard deviation; CR, concentration ratio; GC, gini coefficient. GLO, Global; CN, China; RoW, Rest of the World; RU, Russia; RER, Europe; RNA, Northern America; RME, Middle East.

P	LCIA	Name	FoU	Mean [Rank]		Median [Rank]		SD	CR(80)	GC	h>μ [CR(h>μ)]	
CAUSER	ARD	hard coal//[CN] hard coal mine operation	10,850	0.107	1	0.086	1	0.092	0.038	0.469	4,925	0.79
		petroleum//[RoW] petroleum and gas production, onshore	10,850	0.085	2	0.052	3	0.077	0.025	0.498	4,003	0.73
		petroleum//[RME] petroleum production, onshore	10,850	0.084	3	0.052	4	0.076	0.025	0.498	4,003	0.73
		hard coal//[RNA] hard coal mine operation	10,850	0.071	4	0.051	5	0.093	0.088	0.521	4,327	0.75
		hard coal//[RoW] hard coal mine operation	10,850	0.069	5	0.058	2	0.079	0.082	0.463	5,089	0.77
		petroleum//[RU] petroleum production, onshore	10,850	0.037	6	0.023	6	0.034	0.026	0.499	4,001	0.73
		petroleum//[RoW] petroleum and gas production, offshore	10,850	0.035	7	0.021	8	0.032	0.027	0.499	3,996	0.73
		natural gas, high pressure//[RU] natural gas production	10,850	0.032	8	0.021	9	0.052	0.102	0.552	3,130	0.67
		natural gas, high pressure//[RoW] natural gas production	10,850	0.032	9	0.021	7	0.043	0.073	0.541	3,224	0.66
		natural gas, unprocessed, at extraction//[GLO] natural gas production,	10,850	0.032	10	0.021	10	0.044	0.077	0.544	3,187	0.66
	POFP	electricity, high voltage//[CN] electricity production, hard coal	10,850	0.061	1	0.035	1	0.073	0.056	0.575	3,711	0.76
		diesel, burned in building machine//[GLO] diesel, burned in building machine	10,850	0.060	2	0.019	2	0.114	0.103	0.683	2,630	0.76
		blasting//[RoW] blasting	10,850	0.032	3	0.014	4	0.061	0.121	0.664	2,780	0.74
		coke//[RoW] coking	10,850	0.028	4	0.008	9	0.049	0.089	0.696	2,452	0.78
		transport, freight, sea, transoceanic ship//[GLO] transport, freight, sea,	10,850	0.026	5	0.015	3	0.038	0.092	0.572	3,915	0.77
		diesel, burned in diesel-electric generating set//[GLO] diesel, burned in diesel-	10,850	0.021	6	0.009	6	0.062	0.206	0.669	1,645	0.66
		heat, district or industrial, other than natural gas//[RoW] heat production, at	10,850	0.020	7	0.008	8	0.040	0.124	0.679	2,887	0.77
		natural gas, vented//[GLO] natural gas venting from petroleum/natural gas	10,852	0.017	8	0.011	5	0.031	0.119	0.516	2,644	0.60
		blasting//[RER] blasting	10,850	0.016	9	0.007	13	0.031	0.124	0.665	2,776	0.74
		transport, freight, lorry 16-32 metric ton, EURO3//[RoW] transport, freight, lorry	10,850	0.014	10	0.007	10	0.022	0.103	0.601	3,484	0.74

Table S2-5: Use and contribution patterns of the most important unit processes according to the connector perspective for ARD and POFP. FoU, frequency of use; SD, standard deviation; CR, concentration ratio; GC, gini coefficient. GLO, Global; CN, China; RoW, Rest of the World; RNA, Northern America; RME, Middle East.

P	LCIA	Name	FoU	Mean [Rank]		Median [Rank]		StD	CR(80)	GC	h>μ [CR(h>μ)]	
CONNECTOR	ARD	petroleum//[GLO] market for petroleum	10,850	0.028	1	0.184	1	0.268	0.024	0.498	4,003	0.73
		hard coal//[CN] hard coal mine operation	10,850	0.020	2	0.173	2	0.184	0.038	0.469	4,925	0.79
		petroleum//[RoW] petroleum and gas production, on-shore	10,850	0.016	3	0.107	4	0.158	0.025	0.498	4,003	0.73
		petroleum//[RME] petroleum production, onshore	10,850	0.016	4	0.105	5	0.154	0.025	0.498	4,003	0.73
		hard coal//[RNA] hard coal mine operation	10,850	0.013	5	0.103	6	0.186	0.088	0.521	4,327	0.75
		hard coal//[RoW] hard coal mine operation	10,850	0.013	6	0.117	3	0.160	0.082	0.463	5,089	0.77
		hard coal//[CN] market for hard coal	10,850	0.010	7	0.089	7	0.094	0.038	0.468	4,926	0.79
		diesel//[RoW] market for diesel	10,850	0.010	8	0.039	19	0.153	0.058	0.659	3,042	0.78
		diesel//[RoW] petroleum refinery operation	10,850	0.010	9	0.038	20	0.152	0.059	0.659	3,041	0.78
		natural gas, high pressure//[RoW] market for natural gas, high pressure	10,850	0.009	10	0.067	8	0.134	0.072	0.541	3,229	0.66
	POFP	electricity, high voltage//[CN] electricity production, hard coal	10,850	0.016	1	0.072	1	0.152	0.056	0.575	3,711	0.76
		diesel, burned in building machine//[GLO] diesel, burned in building machine	10,850	0.016	2	0.040	2	0.238	0.103	0.683	2,630	0.76
		diesel, burned in building machine//[GLO] market for diesel, burned in	10,850	0.008	3	0.021	13	0.124	0.103	0.683	2,633	0.76
		blasting//[RoW] blasting	10,850	0.008	4	0.029	7	0.125	0.121	0.664	2,780	0.74
		pig iron//[GLO] market for pig iron	10,850	0.007	5	0.017	20	0.106	0.080	0.704	2,393	0.79
		coke//[RoW] coking	10,850	0.007	6	0.017	19	0.105	0.089	0.696	2,452	0.78
		pig iron//[GLO] pig iron production	10,850	0.007	7	0.016	22	0.104	0.081	0.705	2,391	0.79
		transport, freight, lorry, unspecified//[GLO] market for transport, freight, lorry,	10,850	0.007	8	0.032	4	0.089	0.096	0.596	3,534	0.75
		transport, freight, sea, transoceanic ship//[GLO] transport, freight, sea,	10,850	0.007	9	0.032	5	0.081	0.092	0.572	3,915	0.77
		electricity, high voltage//[CN] market for electricity, high voltage	10,850	0.007	10	0.033	3	0.063	0.052	0.563	3,782	0.75

2.9.3 Evaluation

2.9.3.1 Algorithm to compute the net amount of inputs

Figure S2-15 shows the concept of the algorithm.

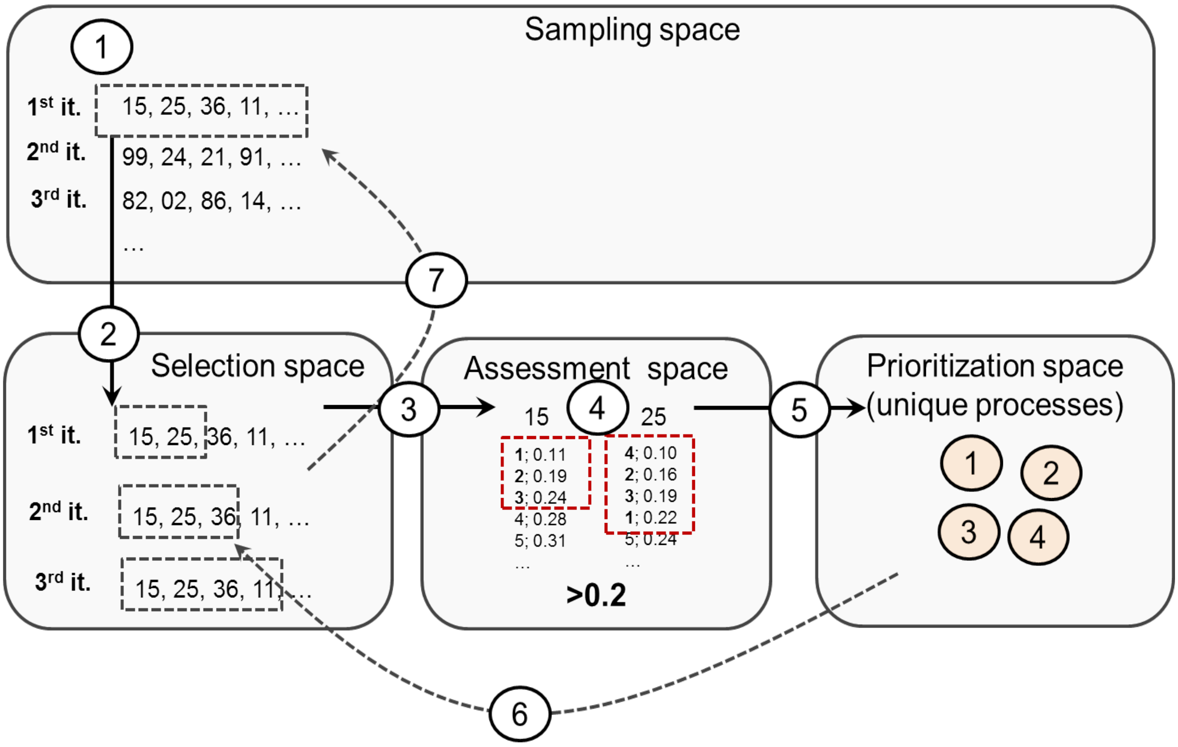


Figure S2-15: Concept schema of the algorithm applied.

The effective amount of inputs required to be reviewed as a function of the considered LCIA indicators and for different thresholds is computed as follows.

1. The algorithm generates 10,000 uniform samples for 112 LCIA indicator indices. Each sample provides the basis for the iterative selection and analysis of the mean input contribution associated with the LCIA indicator indices.
2. The algorithm selects the first sample and passes it to the LCIA selection space.
3. The algorithm identifies the first two LCIA indicator indices in the sample and passes them to an assessment space.
4. For all LCIA indicators in the assessment space, we identify the unit processes required to cross a certain threshold and pass them to the prioritization space.
5. In the prioritization space the algorithm identifies and deletes equal unit processes and counts the effective amount of inputs required per threshold.

6. Each iteration adds one LCIA indicator into the assessment space. The iteration proceeds until all LCIA indicators of a sample have been analyzed.
7. Once all LCIA indicator indices in the selection space have been analyzed, the algorithm proceeds to step 2 and selects the next sample.

Figure S2-16 shows the effective amount of inputs required to be reviewed as a function of the considered LCIA indicators and for different thresholds for the connector perspective.

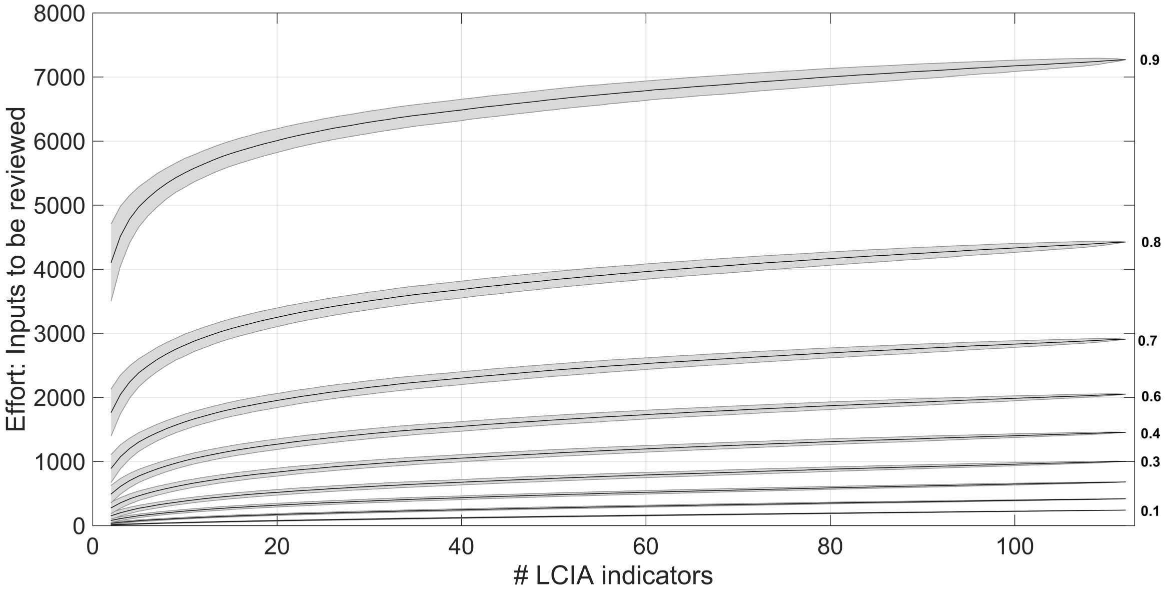


Figure S2-16: Inputs to be reviewed as a function of considered LCIA indicators for different thresholds. Reading example: To review 80% of the contribution across all 112 LCIA indicators we require on average 4,422 connectors. It is the sequence of the LCIA indicator which causes the variation. The black lines represent the mean of 10,000 uniform samples. The gray lines represent the corresponding standard deviation.

Table S2-6 shows the “modern” LCIA indicators used for the computation.

Table S2-6: “Modern” LCIA indicators used for the computation of the amount of inputs required to review for different thresholds.

No.	LCIA indicator
1	('CML 2001', 'terrestrial ecotoxicity', 'TAETP infinite')
2	('CML 2001', 'human toxicity', 'HTP infinite')
3	('CML 2001', 'climate change', 'GWP 100a')
4	('CML 2001', 'land use', 'competition')
5	('CML 2001', 'acidification potential', 'generic')
6	('CML 2001', 'freshwater sediment ecotoxicity', 'FSETP infinite')
7	('CML 2001', 'eutrophication potential', 'generic')
8	('CML 2001', 'freshwater aquatic ecotoxicity', 'FAETP infinite')
9	('CML 2001', 'marine aquatic ecotoxicity', 'MAETP infinite')
10	('CML 2001', 'ionising radiation', 'ionising radiation')

11	('CML 2001', 'photochemical oxidation (summer smog)', 'MIR')
12	('CML 2001', 'marine sediment ecotoxicity', 'MSETP infinite')
13	('CML 2001', 'stratospheric ozone depletion', 'ODP steady state')
14	('CML 2001', 'resources', 'depletion of abiotic resources')
15	('cumulative energy demand', 'fossil', 'non-renewable energy resources, fossil')
16	('cumulative energy demand', 'nuclear', 'non-renewable energy resources, nuclear')
17	('cumulative energy demand', 'biomass', 'renewable energy resources, biomass')
18	('cumulative energy demand', 'water', 'renewable energy resources, potential (in barrage water), converted')
19	('cumulative energy demand', 'geothermal', 'renewable energy resources, geothermal, converted')
20	('cumulative energy demand', 'solar', 'renewable energy resources, solar, converted')
21	('cumulative energy demand', 'wind', 'renewable energy resources, kinetic (in wind), converted')
22	('cumulative energy demand', 'primary forest', 'non-renewable energy resources, primary forest')
23	('ecological scarcity 2013', 'total', 'Radioactive substances into air')
24	('ecological scarcity 2013', 'total', 'POP into water')
25	('ecological scarcity 2013', 'total', 'Energy resources')
26	('ecological scarcity 2013', 'total', 'Pesticides into soil')
27	('ecological scarcity 2013', 'total', 'Heavy metals into water')
28	('ecological scarcity 2013', 'total', 'Heavy metals into air')
29	('ecological scarcity 2013', 'total', 'Water pollutants')
30	('ecological scarcity 2013', 'total', 'Mineral resources')
31	('ecological scarcity 2013', 'total', 'Non radioactive waste to deposit')
32	('ecological scarcity 2013', 'total', 'Land use')
33	('ecological scarcity 2013', 'total', 'Global warming')
34	('ecological scarcity 2013', 'total', 'Water resources')
35	('ecological scarcity 2013', 'total', 'Radioactive substances into water')
36	('ecological scarcity 2013', 'total', 'Carcinogenic substances into air')
37	('ecological scarcity 2013', 'total', 'Main air pollutants and PM')
38	('ecological scarcity 2013', 'total', 'Heavy metals into soil')
39	('ecological scarcity 2013', 'total', 'Radioactive waste to deposit')
40	('ecological scarcity 2013', 'total', 'Ozone layer depletion')
41	('ecological scarcity 2013', 'total', 'total')
42	('IMPACT 2002+ (Endpoint)', 'climate change', 'climate change')
43	('IMPACT 2002+ (Endpoint)', 'ecosystem quality', 'terrestrial ecotoxicity')
44	('IMPACT 2002+ (Endpoint)', 'ecosystem quality', 'terrestrial acidification & nutrification')
45	('IMPACT 2002+ (Endpoint)', 'ecosystem quality', 'aquatic ecotoxicity')
46	('IMPACT 2002+ (Endpoint)', 'ecosystem quality', 'land occupation')
47	('IMPACT 2002+ (Endpoint)', 'human health', 'ionising radiation')
48	('IMPACT 2002+ (Endpoint)', 'human health', 'photochemical oxidation')
49	('IMPACT 2002+ (Endpoint)', 'human health', 'human toxicity')
50	('IMPACT 2002+ (Endpoint)', 'human health', 'respiratory effects (inorganics)')
51	('IMPACT 2002+ (Endpoint)', 'human health', 'ozone layer depletion')
52	('IMPACT 2002+ (Endpoint)', 'resources', 'mineral extraction')

53	('IMPACT 2002+ (Endpoint)', 'resources', 'non-renewable energy')
54	('IMPACT 2002+ (Endpoint)', 'climate change', 'total')
55	('IMPACT 2002+ (Endpoint)', 'resources', 'total')
56	('IMPACT 2002+ (Endpoint)', 'human health', 'total')
57	('IMPACT 2002+ (Endpoint)', 'ecosystem quality', 'total')
58	('IMPACT 2002+ (Midpoint)', 'ecosystem quality', 'aquatic acidification')
59	('IMPACT 2002+ (Midpoint)', 'ecosystem quality', 'aquatic eutrophication')
60	('IPCC 2013', 'climate change', 'GWP 100a')
61	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'terrestrial ecotoxicity')
62	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'natural land transformation')
63	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'urban land occupation')
64	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'agricultural land occupation')
65	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'climate change, ecosystems')
66	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'freshwater eutrophication')
67	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'terrestrial acidification')
68	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'freshwater ecotoxicity')
69	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'marine ecotoxicity')
70	('ReCiPe Endpoint (H,A)', 'human health', 'human toxicity')
71	('ReCiPe Endpoint (H,A)', 'human health', 'ionizing radiation')
72	('ReCiPe Endpoint (H,A)', 'human health', 'particulate matter formation')
73	('ReCiPe Endpoint (H,A)', 'human health', 'climate change, human health')
74	('ReCiPe Endpoint (H,A)', 'human health', 'photochemical oxidant formation')
75	('ReCiPe Endpoint (H,A)', 'human health', 'ozone depletion')
76	('ReCiPe Endpoint (H,A)', 'resources', 'fossil depletion')
77	('ReCiPe Endpoint (H,A)', 'resources', 'metal depletion')
78	('ReCiPe Endpoint (H,A)', 'total', 'total')
79	('ReCiPe Endpoint (H,A)', 'ecosystem quality', 'total')
80	('ReCiPe Endpoint (H,A)', 'resources', 'total')
81	('ReCiPe Endpoint (H,A)', 'human health', 'total')
82	('ReCiPe Midpoint (H)', 'terrestrial ecotoxicity', 'TETPinf')
83	('ReCiPe Midpoint (H)', 'natural land transformation', 'NLTP')
84	('ReCiPe Midpoint (H)', 'photochemical oxidant formation', 'POFP')
85	('ReCiPe Midpoint (H)', 'human toxicity', 'HTPinf')
86	('ReCiPe Midpoint (H)', 'marine eutrophication', 'MEP')
87	('ReCiPe Midpoint (H)', 'climate change', 'GWP100')
88	('ReCiPe Midpoint (H)', 'particulate matter formation', 'PMFP')
89	('ReCiPe Midpoint (H)', 'agricultural land occupation', 'ALOP')
90	('ReCiPe Midpoint (H)', 'freshwater eutrophication', 'FEP')
91	('ReCiPe Midpoint (H)', 'metal depletion', 'MDP')
92	('ReCiPe Midpoint (H)', 'terrestrial acidification', 'TAP100')
93	('ReCiPe Midpoint (H)', 'water depletion', 'WDP')
94	('ReCiPe Midpoint (H)', 'urban land occupation', 'ULOP')
95	('ReCiPe Midpoint (H)', 'ionising radiation', 'IRP_HE')
96	('ReCiPe Midpoint (H)', 'fossil depletion', 'FDP')

97 ('ReCiPe Midpoint (H)', 'freshwater ecotoxicity', 'FETPinf')
98 ('ReCiPe Midpoint (H)', 'marine ecotoxicity', 'METPinf')
99 ('ReCiPe Midpoint (H)', 'ozone depletion', 'ODPinf')
100 ('TRACI', 'human health', 'non-carcinogenics')
101 ('TRACI', 'human health', 'respiratory effects, average')
102 ('TRACI', 'human health', 'carcinogenics')
103 ('TRACI', 'environmental impact', 'eutrophication')
104 ('TRACI', 'environmental impact', 'global warming')
105 ('TRACI', 'environmental impact', 'ecotoxicity')
106 ('TRACI', 'environmental impact', 'photochemical oxidation')
107 ('TRACI', 'environmental impact', 'acidification')
108 ('TRACI', 'environmental impact', 'ozone depletion')
109 ('USEtox', 'human toxicity', 'non-carcinogenic')
110 ('USEtox', 'human toxicity', 'carcinogenic')
111 ('USEtox', 'ecotoxicity', 'total')
112 ('USEtox', 'human toxicity', 'total')

2.9.4 Discussion

2.9.4.1 Sector relevance

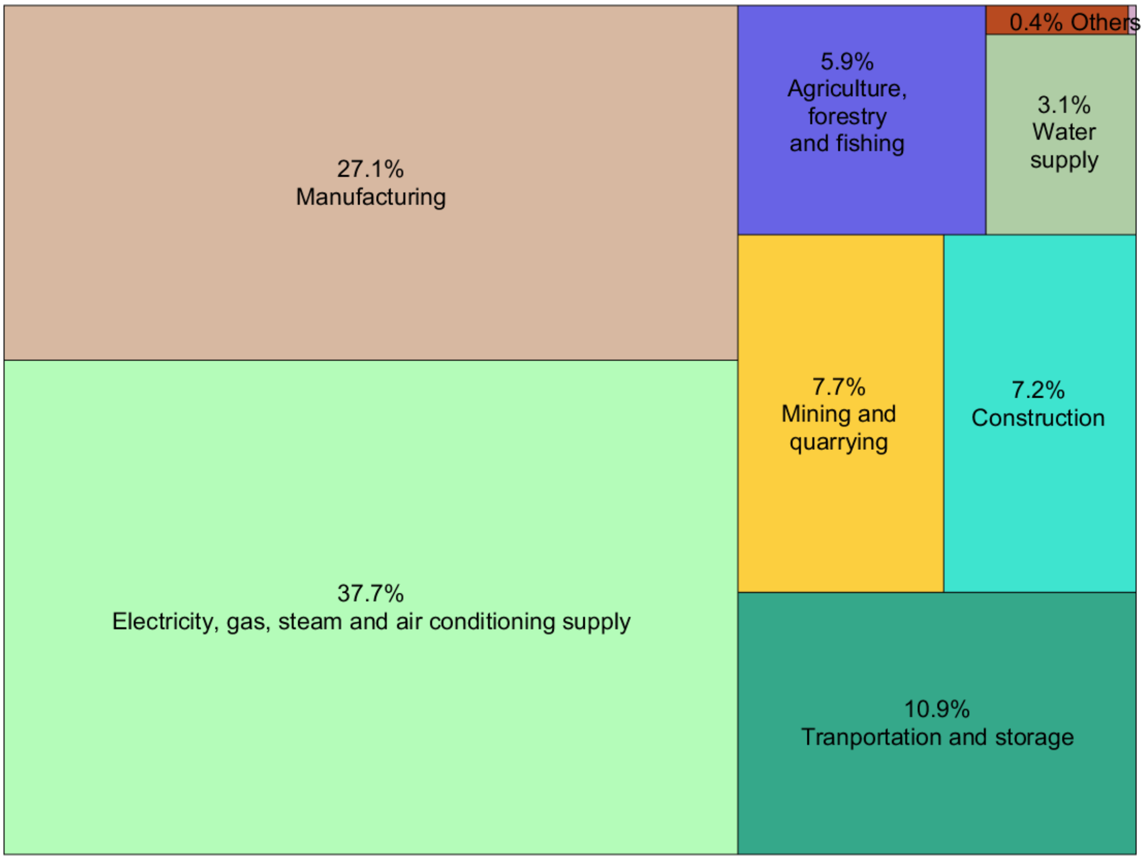


Figure S2-17: Tree plot showing the relative contribution of the most important sector –relative sector importance—according to ISIC rev. 4 classifications throughout the entire ecoinvent database for the POFP causer perspective. The name of the sectors with a contribution <0.5% are expressed as "Others". The contribution-based view could be contrasted with the sectorial distribution of datasets. Similar analysis can be done to analyze the mean relevance of infrastructure processes.

3 ARTICLE II: CONTRIBUTION-BASED PRIORITIZATION OF LCI DATABASE IMPROVEMENTS. THE MOST IMPORTANT UNIT PROCESSES IN ECOINVENT

Juergen Reinhard[°*], Christopher L. Mutel[◊], Gregor Wernet[#], Rainer Zah[★], Lorenz M. Hilty[°*],

°University of Zurich, Department of Informatics, Informatics and Sustainability Research Group, Binzmuehlestrasse 14, 8050 Zurich, Switzerland. Email: juergen.reinhard@uzh.ch

◊ Laboratory for Energy Systems Analysis, Paul Scherrer Institute, 5232 Villingen PSI, Switzerland

#ecoinvent, Technoparkstrasse 1, 8057 Zürich, Switzerland

★Quantis Serl, Reitergasse 11, 8600 Zurich, Switzerland

**Empa Materials Science and Technology, Technology and Society Laboratory, Ueberlandstrasse 129, 8600 Duebendorf, Switzerland. Tel. +41 58 765 4207*

Status: *Submitted to the International Journal of Life Cycle Assessment.*

Abstract

Purpose:

Improving the quality and quantity of unit process datasets in Life Cycle Inventory (LCI) databases affects every LCA they are used in. However, improvements in data quality and quantity are so far rather directed by the external supply of data and situation-driven requirements instead of systematic choices guided by structural dependencies in the data. Overall, the impact of current data updates on the quality of the LCI database remains unclear and maintenance efforts might be inefficient. This paper analyzes how a contribution-based prioritization approach can direct LCI update efforts to datasets of key importance.

Methods:

A contribution-based prioritization method has been applied to version 3.1 of the ecoinvent database. We identified the importance of unit processes on the basis of their relative contributions throughout each product system with respect to a broad range of Life Cycle Impact Assessment (LCIA) indicators. A novel consolidation algorithm enabled the ranking of unit processes according to their impact on the LCIA results. Finally, we identified the most relevant LCI datasets for different sectors and geographies.

Results and discussion:

The study shows that a relatively large proportion of the overall database quality is dependent on a small set of key processes. Processes related to electricity generation, waste treatment activities and energy carrier provision (petroleum and hard coal) consistently cause large environmental impacts in all product systems. Overall, 300 datasets are causing 60% of the environmental impacts across all LCIA indicators, while only 3 datasets are causing 11% of all climate change impacts. In addition, our analysis highlights

the presence and importance of central hubs, i.e., sensitive intersections in the database network, whose modification can affect a large proportion of database quality.

Conclusion:

Our study suggests that contribution-based prioritization offers important insights for the systematic and effective improvement of LCI databases. The presented list of key processes in ecoinvent version 3.1 adds a new perspective to database improvements as it allows the allocation of available resources according to the structural dependencies in the data.

Keywords

Life Cycle Inventory database management; Prioritization; Meta analysis; Contribution analysis; Life Cycle Assessment

Electronic supplementary material

This article contains supplementary material in two file formats: Supplementary Information 1 (section 3.7) is a file which complements the article with additional figures and discussions, and SI2 in Microsoft Excel (.xlsx) format, a file provided for dynamic exploration of the data processed and discussed in the article.

3.1 Introduction

Life Cycle Assessment is a technique for the comprehensive, quantitative assessment of the environmental impacts of products[39] throughout their entire value chain (Finnveden et al., 2009). This involves the mapping of complex globalized networks consisting of thousands of interlinked human activities (Hellweg and Milà i Canals, 2014). Tracing and measuring the exchange flows of and between these activities is realized on the basis of nodes called unit processes. A unit process represents one specific activity or a group of activities allocated to one unique output and records (i) the *exchanges with environment*, i.e., the input of natural resources and output of emissions (ii) the *intermediate exchanges* from and to the technosphere, i.e., the input of usable energy and raw materials and the output of products and waste (Reinhard et al., 2016).

A typical value chain covers thousands of unit processes, each of which needs to be described with exchange flow values (Bourgault et al., 2012). This information cannot usually be gathered as primary data within a specific project due to the high cost that would be involved in data collection (Reinhard et

[39] The term product includes both goods and services.

al., 2016). It is therefore common practice to focus data collection efforts on selected activities that reflect the space for action—these activities are together called the foreground system—and to use generic data from Life Cycle Inventory (LCI) databases[40] to model the remaining activities, called the background system (Bourgault et al., 2012; Reinhard et al., 2016; Tillman, 2000). Even when 100 processes are modeled with primary data, the foreground system would still not exceed 5% of the entire product system as it typically involves more than several thousand unit processes (Steubing et al., 2016). Bearing this in mind, background data from LCI databases can be considered the backbone of any LCA study (Reinhard et al. submitted). Their unit process data form the basic building blocks required by all LCA applications (Sonnemann and Vigon, 2011).

However, unit process datasets are subject to uncertainties. The exchange flow data required for the accurate compilation of a unit processes can be *unavailable, wrong, or unreliable* (Ciroth et al., 2013; Heijungs and Huijbregts, 2004). As unit processes typically represent average conditions of a whole country, a given time period and different instances of real processes, *natural variability* is always present (Huijbregts, 1998). Both cases affect the overall accuracy of the unit process (Sonnemann and Vigon, 2011). Consequently, LCI databases like ecoinvent are under continuous extension and improvement.

However, improvements are becoming progressively difficult due to the increasing numbers of datasets stored in existing databases. For example, version 3.1 of the ecoinvent database contains about 10,000 unit process datasets. Assuming that updating one unit process would, on average, require one working day, then updating the entire database would roughly require the continuous work of fifty-five workers for one year. Capacities for such extensive improvement efforts are typically not available. In addition, the desired quality and quantity of databases is a constantly moving target. It is the "continuing evolution in consumer preferences, market and industry imperatives, and public policy which forces continuous development and improvement of datasets and methodologies for LCA to meet these needs" (Sonnemann and Vigon, 2011, p. 98). Compliance with these developments typically ties up a lot of the available workforce. Consequently, the focus of inventory efforts is strongly guided by external requirements and data availability instead of inherent structural importance of unit process datasets.

The goal of this paper is therefor to rank the unit processes of ecoinvent 3.1 considering their importance across 19 LCIA indicators according to the following steps:

- We apply the prioritization method from Reinhard et al. (2016) to the version 3.1 of the ecoinvent database. We first focus on three selected LCIA indicators in order to establish a basic understanding and to highlight important characteristics of the approach.

[40] Our definition of LCI database follows the definition of the Shonan Guidance Principles (Sonnemann and Vigon, 2011).

- We identify and present the most important unit processes according to a set of 19 selected LCIA indicators using a newly developed consolidation algorithm. The algorithm calculates the overall rank of unit processes considering their importance across all LCIA indicators. We use the final prioritization list to identify the sectors and locations of particular importance.
- We discuss and analyze possible reasons for the presence of systemic datasets and associated insights. We also define the limitations of the approach and related future work.

3.2 Method

3.2.1 Contribution-based prioritization

3.2.1.1 Application to LCI databases

The application of contribution analysis (CA) is quite common in LCA and is implemented in all of the commercially available software tools (Ciroth et al., 2013; Ebner, 2013; Goedkoop and Oele, 2004). CA focuses on the disaggregation of the aggregated results in order to identify the elements that make the highest contribution (Heijungs and Kleijn, 2001). Reinhard et al. (2016) has formalized CA for the application on full LCI databases. Figure 3-1 shows the workflow of the method on the basis of a simplified LCI database example consisting of three products (A, B & C), one elementary flow (CO_2) and one corresponding characterization factor.

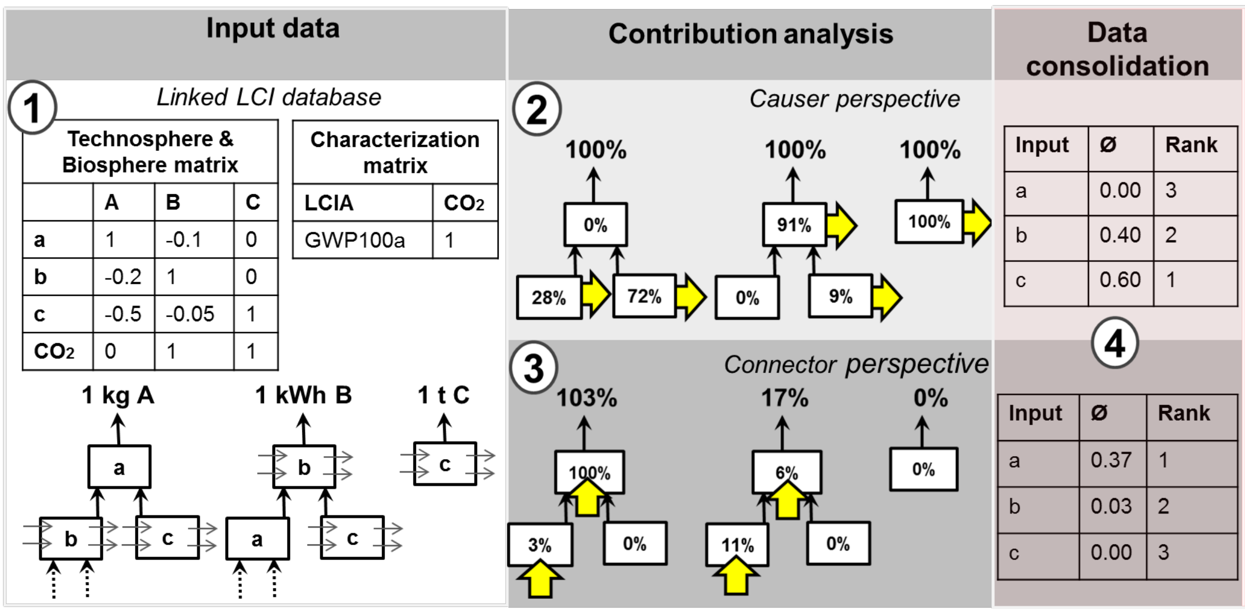


Figure 3-1: Workflow of the prioritization method explained on the basis of a simplified example. (1) The product systems are shown in matrix notation and as tree graphs (2 & 3). We calculate the contributions of the processes in two ways in order to highlight their causing and their connecting characteristics. (4) We consolidate and rank the contributions associated with a particular process (a, b or c) based on Ø; the arithmetic mean of a process contributions throughout all product systems, e.g., [1/3 *(0.28 +0.91+ 0] for B in the causer perspective. Source: authors.

First, using the linked LCI database (step 1 in Figure 3-1) in matrix notation and one row from the characterization matrix we calculate the environmental impact of every product system in the database for a unitary demand, i.e. for one unit of each product. Next, using the formulas specified in Table S3-4 (section 3.7.1.1), we calculate the relative LCIA contributions of each process (a, b & c) throughout all product systems according to two different perspectives; the *causer perspective* (step 2) and the *connector or network perspective* (step 3). These perspectives represent "complementary lenses for our goal of identifying important processes throughout the database" (Reinhard et al., 2016).

We first calculate the LCIA contributions of the *causing elements* of each unit process (step 2). Causing elements are the direct exchanges with the environment that cause environmental impacts. For example, process *b* and *c* causes 1 kg CO_2 emission, respectively. Therefore, both processes are related to a relative contribution throughout all product systems in the causer perspective. Vice versa, process *a* causes no exchange to the environment and therefore has a contribution of zero. The causer perspective helps us to pinpoint the processes with consistently large contributions in terms of environmental interventions.

We then calculate the LCIA contributions of the *connecting elements* of each unit process (step 3), that is, the intermediate flows from other processes in the technosphere. From this perspective, the environmental impact of process *a* is exclusively determined by the environmental impact associated with its intermediate exchanges, i.e. *b* and *c*. As *c* requires no intermediate exchanges it has a relative contribution of zero throughout all product system in this perspective. The connector perspective helps to pinpoint the unit processes that consistently link to large upstream contributions. It results in double counting because the upstream contributions counted in one process include shares of the contribution already accounted for in the preceding process. Therefore, the relative contributions do not add up to 100% of the impact.

Finally, we use the arithmetic mean (Ø) to consolidate (step 4) the contribution patterns of the processes. We use the Ø because we are interested in typical values representing the "real" balance point of the set of contributions (Bulmer, 1979). This gives a useful indication of the average importance of one process throughout the full database. The Ø of a particular unit process in the causer perspective expresses the relative average contribution caused by its elementary flows throughout all product systems in the database. The Ø of a particular unit process in the connector perspective expresses the amount of relative contribution which, on average, is transmitted by its intermediate inputs throughout all product systems in the database. Note that we use the modulus for the calculation of the Ø (see formula 13 in Table S3-4). This avoids distortions due to negative contributions and ensures that the actual size of the contributions – the relevant information for our purpose – is correctly accounted for. Negative contributions can result from negative characterization factors and/or negative elementary flows (e.g. an uptake of carbon dioxide) in the biosphere matrix. This workflow has to be repeatedly executed for every LCIA indicator of interest.

3.2.1.2 Implementation

Version 3.1 of the ecoinvent database offers three system models. A system model consists of a predefined set of rules for the transformation of unlinked multi-output activities into interlinked, single product processes. We work with the system model "Allocation, cut-off by classification"[41] (cut-off). The cut-off system model results in 11,304 unit processes, i.e. a square technosphere matrix of the size 11,304 x 11,304 (see Table S3-6 in section 3.7.2.1).

We first loaded the linked database model and 19 pre-selected LCIA indicators into MATLAB and applied formula 5-16 (Table S3-4) to compute **Rd** and **Rup** for each LCIA indicator. We selected LCIA indicators (Table S3-5) based on their scientific quality[42], their availability[43], and their assumed relevance in practice. The processing of all LCIA indicators results in roughly 4.8 billion potential data points (19 LCA indicators, two matrixes (one for each perspective) with a size of 11,304 x 11,304). Figure S1 in the SI illustrates the overall relative contribution matrixes for one LCIA indicator.

Next, we calculated the mean contribution of each process listed in **Rd** and **Rup** (Table S3-4, formula 17) according to all LCIA indicators. Likewise, we counted the frequency of use (FoU) of each process throughout all product systems by replacing all non-zero contributions with one and summing up the result.

Subsequently, we ranked the processes according to their mean process contributions in descending order. Due to the limitations in space LCIA indicator specific results are only presented for CC, Etox, and ReCiPe. We aimed to highlight important characteristics of the most important processes according to these different LCIA indicators.

Finally, we used a novel consolidation algorithm to calculate an overall rank of unit processes considering their importance across all 19 LCIA indicators.

3.2.2 Prioritizing across many LCIA indicators

Prioritization across 19 LCIA indicators is difficult to interpret for each LCIA indicators in isolation. Therefore, we developed a consolidation algorithm based on the Lorenz curve (Duclos and Araar, 2006)

[41] The system model is based on the Cut-off approach where primary (first) production of materials is always allocated to the primary user of a material. Furthermore, a primary producer of a recyclable material does not receive any credit for its provision. Therefore, recyclable materials are available burden-free to recycling processes, and secondary (recycled) materials bear only the impacts of the recycling processes (Wernet et al., 2016).

[42] Scientific quality is determined according to the recommendations in the ILCD-Handbook (EC-JRC, 2011).

[43] Though the ecoinvent center provides the aggregated results of 692 LCIA indicators, it does not cover all LCIA indicators recommended by the ILCD-Handbook.

that allows the calculation of an overall rank of unit processes considering their importance across all LCIA indicators.

We first construct a mirror image of the Lorenz curve (MLC) for each LCIA indicator and perspective. The MLC indicates the cumulative percentage of total mean contribution held by a cumulative proportion of the unit processes. We start with the process whose mean contribution throughout the database is the largest and proceed by adding all other mean input contributions in descending order (Reinhard et al., 2016).

Figure 3-2 shows how the algorithm would prioritize across the cumulated contribution of four processes, two LCIA indicators and three thresholds. The actual data, i.e., the descending ordered and accumulated list of mean process contributions for all unit processes and all LCIA indicators can be examined in SI2.

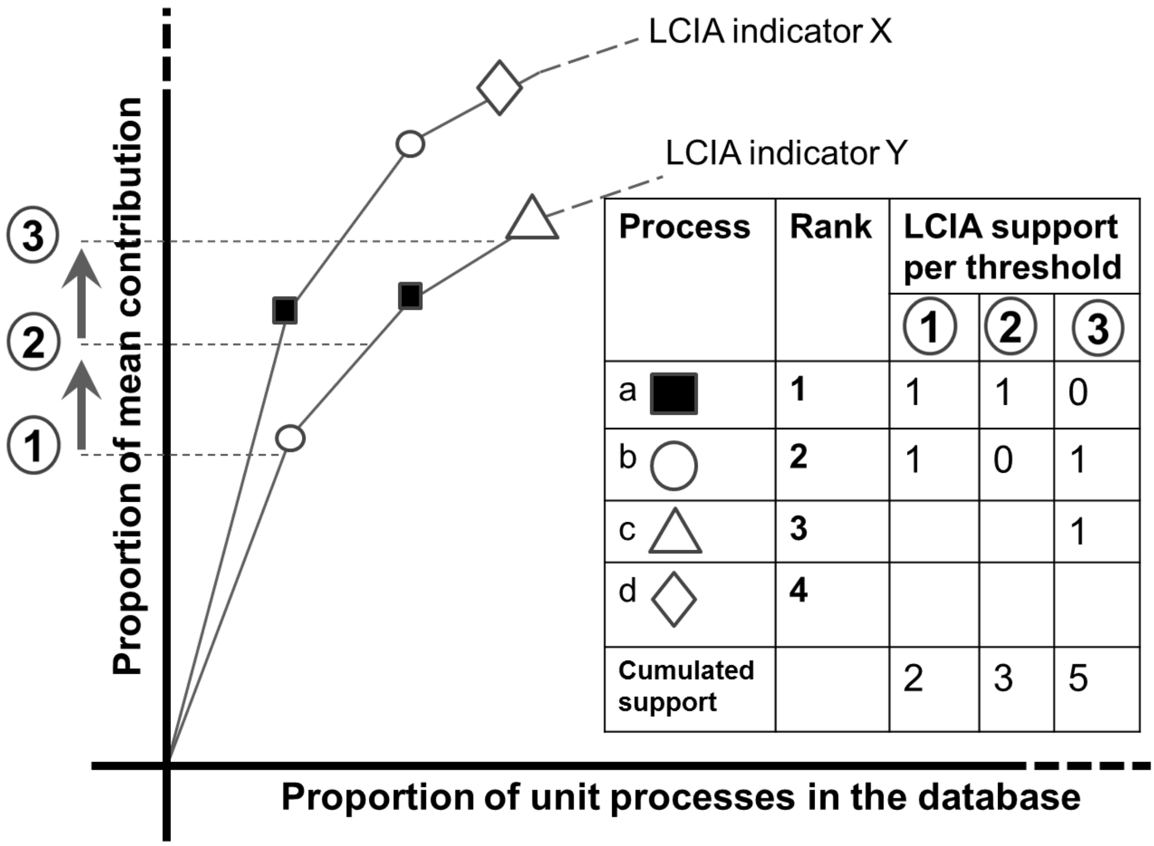


Figure 3-2: Visual illustration of the prioritization across the cumulated contribution of four processes, two LCIA indicators and three thresholds.

For example, the minimum amount of processes required to exceed the first threshold amounts to one for both LCIA indicator, i.e. we require process *a* (LCIA indicator X) and process *b* (LCIA indictor Y). That is, two processes generate an overall LCIA support of two, i.e., there are no duplicates in the prioritization space. To exceed the second threshold, we require only process *a* from LCIA indicator Y as process *a* of LCIA indicator X and process *b* of LCIA indicator Y is already accounted for in the first threshold. That is, two processes generate a cumulated LCIA support of three, i.e., there is one duplicate (process *a*, the square) in the prioritization space. Further, to exceed the third threshold, we require process *b* of LCIA indicator X as well as process *c* from LCIA indicator Y. Hence, three processes generate an overall LCIA support of five.

3.3 Results

3.3.1 Prioritization according to selected LCIA indicators

Table 3-1 and Table 3-2 show the frequency of use (FoU) and mean process contribution for the seven largest processes (Ø); once for the causer and once for the connector perspective, respectively. The results are sorted in descending order according to the Ø. The last column, "Ø cumulated", shows how much of the overall contribution a certain amount of processes accumulates. The full list covering all processes and all LCIA indicators is available in the excel file, i.e., SI2.

Table 3-1: Excerpt of the summary table for the seven largest causers for three selected LCIA indicators. The full list can be viewed in the SI2. Ecoinvent uses geographical shortcuts (also shown in SI2): CN = China; RoW= Rest-of-the-World; GLO = Global; RME=Middle East; RNA=Northern America. RoW has been introduced to cover all location where a local process is not yet available. Consequently, the spatial scope of RoW varies. Currently, it can represents more than 100 different locations (Wernet et al., submitted).

LCIA	**Rank**	**Name [product //[geographical location] activity]**	**FoU**	**Ø**	**Ø cumulated**
CC	1	electricity, high voltage//[CN] electricity production, hard coal	10'850	0.051	0.05
	2	hard coal//[CN] hard coal mine operation	10'850	0.032	0.08
	3	clinker//[RoW] clinker production	10'850	0.031	0.11
	4	heat, district or industrial, other than natural gas//[RoW] heat production, at hard coal industrial furnace 1-10MW	10'850	0.030	0.14
	5	diesel, burned in building machine//[GLO] diesel, burned in building machine	10'850	0.025	0.17
	6	pig iron//[GLO] pig iron production	10'850	0.024	0.19
	7	electricity, high voltage//[IN] electricity production, hard coal	10'850	0.016	**0.21**
ReCiPe	1	hard coal//[CN] hard coal mine operation	10'850	0.030	0.03
	2	sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	10'852	0.030	0.06
	3	petroleum//[RoW] petroleum and gas production, on-shore	10'850	0.023	0.08
	4	electricity, high voltage//[CN] electricity production, hard coal	10'850	0.023	0.11
	5	petroleum//[RME] petroleum production, onshore	10'850	0.023	0.13
	6	hard coal//[RoW] hard coal mine operation	10'850	0.017	0.15
	7	hard coal//[RNA] hard coal mine operation	10'850	0.015	**0.16**
Etox	1	sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	10'852	0.215	0.21
	2	scrap steel//[RoW] treatment of scrap steel, municipal incineration	10'850	0.144	0.36
	3	scrap copper//[RoW] treatment of scrap copper, municipal incineration	10'850	0.139	0.50
	4	spoil from hard coal mining//[GLO] treatment of spoil from hard coal mining, in surface landfill	10'852	0.077	0.57
	5	spoil from lignite mining//[GLO] treatment of spoil from lignite mining, in surface landfill	10'852	0.045	0.62
	6	slag, unalloyed electric arc furnace steel//[RoW] treatment of slag, unalloyed electric arc furnace steel, residual material landfill	10'850	0.026	0.65
	7	natural gas, unprocessed, at extraction//[GLO] natural gas production, unprocessed, at extraction	10'850	0.015	**0.66**

Table 3-2: Excerpt of the summary table for the seven largest connectors for three selected LCIA indicators. ØN represent the normalized mean contribution, i.e. the Ø of a particular process relative to the contribution of all processes. The full list can be viewed in the SI2. Used geographical shortcuts: CN = China; RoW= Rest-of-the-World; GLO = Global; RME=Middle East; RAS=Russia.

LCIA	Rank	Name [product //[geographical location] activity]	FoU	ØN	ØN cumulated	Ø	Ø cumulated
CC	1	electricity, high voltage//[CN] electricity production, hard coal	10'850	0.015	0.02	0.117	0.12
	2	pig iron//[GLO] pig iron production	10'850	0.009	0.02	0.071	0.19
	3	hard coal//[CN] hard coal mine operation	10'850	0.009	0.03	0.065	0.25
	4	clinker//[RoW] clinker production	10'850	0.009	0.04	0.065	0.32
	5	heat, district or industrial, other than natural gas//[RoW] heat production, at hard coal industrial furnace 1-10MW	10'850	0.009	0.05	0.064	0.38
	6	electricity, high voltage//[CN] market for electricity, high voltage	10'850	0.007	0.06	0.055	0.44
	7	diesel, burned in building machine//[GLO] diesel, burned in building machine	10'850	0.007	**0.07**	0.055	**0.49**
ReCiPe	1	petroleum//[GLO] market for petroleum	10'850	0.012	0.01	0.092	0.09
	2	electricity, high voltage//[CN] electricity production, hard coal	10'850	0.008	0.02	0.065	0.16
	3	hard coal//[CN] hard coal mine operation	10'850	0.008	0.03	0.063	0.22
	4	sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	10'852	0.007	0.04	0.058	0.28
	5	copper//[GLO] market for copper	10'850	0.006	0.04	0.049	0.33
	6	petroleum//[RoW] petroleum and gas production, on-shore	10'850	0.006	0.05	0.048	0.37
	7	petroleum//[RME] petroleum production, onshore	10'850	0.006	**0.05**	0.047	**0.42**
Etox	1	sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	10'852	0.046	0.05	0.423	0.42
	2	scrap steel//[RoW] treatment of scrap steel, municipal incineration	10'850	0.031	0.08	0.283	0.71
	3	scrap copper//[RoW] treatment of scrap copper, municipal incineration	10'850	0.030	0.11	0.274	0.98
	4	sulfidic tailing, off-site//[GLO] market for sulfidic tailing, off-site	10'851	0.023	0.13	0.211	1.19
	5	spoil from hard coal mining//[GLO] treatment of spoil from hard coal mining, in surface landfill	10'852	0.016	0.14	0.151	1.34
	6	scrap steel//[GLO] market for scrap steel	10'850	0.015	0.16	0.141	1.48
	7	scrap copper//[GLO] market for scrap copper	10'850	0.015	**0.17**	0.137	**1.62**

In order to illustrate the individual contributions summarized by the Ø in Table 3-1 and Table 3-2, Figure 3-3 visualizes the detailed contribution patterns of the process with the highest contribution for the CC LCIA indicator, "electricity, high voltage//[CN] electricity production, hard coal", once for the connector and once for the causer perspective, respectively.

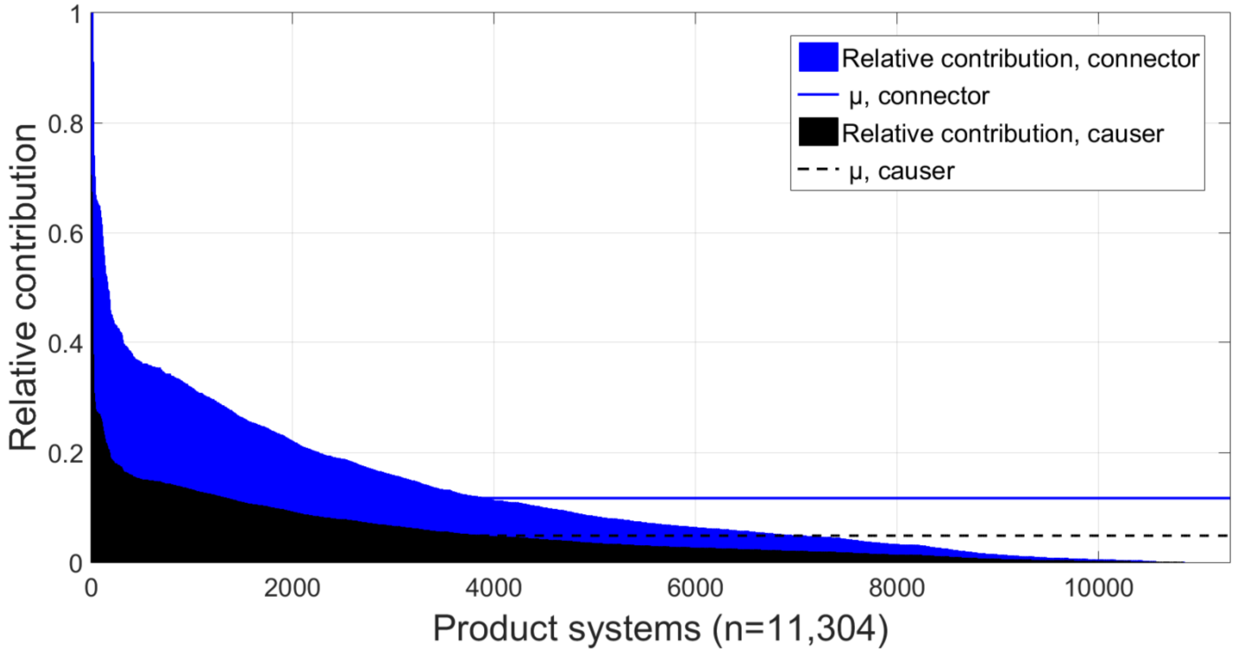


Figure 3-3: Bar chart showing the relative contribution of "electricity, high voltage//electricity production, hard coal [CN]" throughout all 11,304 product systems for the LCIA indicator climate change (CC). In order to improve readability the individual contributions are sorted in descending order.

Figure 3-3 in combination with Table 3-1 and Table 3-2 reveals several notable insights.

First, as illustrated by Figure 3-3, the process is practically used by the entire database, i.e. throughout 10,850[44] (or 96%) of the available product systems (see the FoU indicator in Table 3-1 and Table 3-2, respectively). It is noteworthy, that this holds true for all processes which exceed a mean contribution of 0.5% (see Figure S3-9 and Figure S3-11 in section 3.7.2.2). This means, that an improvement of the largest processes will practically affect all product systems in the database.

Second, the largest connectors typically transmit more environmental impact (contribution) than produced by the largest causers. For example, as shown by Figure 3-3 the connecting elements of electricity production, hard coal [CN] relocate on average 2.4 times more environmental impacts (11.7%) than directly produced by the causing elements (5.1%) of this process. As shown by Table 3-1 and Table 3-2, the Ø of the largest processes in the connector perspective are always larger than the Ø of the largest processes in the causer perspective. This holds true for all processes and all LCIA indicators (see Figure

[44] As shown in Table S1 in SI1, around 450 inputs have no input from the technosphere and no elementary flows of relevance for the selected LCIA indicator.

S3-13 in section 3.7.2.5 for a visual illustration). That is, the accuracy of the intermediate exchanges of the major connectors has a large influence on the quality and the results of all product systems in the database.

Third, the inequalities among the mean process contributions are remarkable. That is, a few processes have consistently large contributions throughout a lot of product system. With regard to the causer perspective, the seven largest causers already accumulate 16% (ReCiPe), 21% (CC) and 66% (Etox) of the overall contribution. Due to the mentioned size differences between the major causers and connectors, the seven largest connectors accumulate 42% (ReCiPE), 49% (CC) and 169% (Etox) of the contribution. The rescaling to ØN, which expresses the relative proportion in relation to the contribution transmitted via all connectors, translates this into 5% (ReCiPe), 7% (CC) and 17% (Etox), respectively (column "CumulatedN" in Table 3-2). The rescaling has a strong influence because almost all processes in the database include connecting elements that link to a contribution. For example, only 3,501 unit processes include an elementary flow relevant for the LCIA indicator CC but 10,869 processes link to an intermediate input with a CC impact (see Figure S3-8 and Figure S3-10 in section 3.7.2.2).

Fourth, we can notice a certain degree of correlation between the causer and the connector perspective, i.e., one and the same unit process can be important according to both perspectives. For example, "electricity, high voltage//[CN] electricity production, hard coal " and "hard coal//[CN] hard coal mine operation" are consistently important according to both perspectives. At the same time, however, the most important connector, "petroleum//[GLO] market for petroleum" according to ReCiPe has no contribution in the causer perspective. This holds true for many other connectors as well and becomes apparent when looking at the correlation between both perspectives (see Figure S3-12 in section 3.7.2.4).

Fifth, we can spot some moderate correlation across LCIA indicators, i.e., different LCIA indicators will point to the same unit processes. For example, "clinker//[RoW] clinker production" and "sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site" are consistently important across both ReCiPe and CC. At the same time, however, different LCIA indicators have, due to their different foci, quite different inventory support[45], and thus prioritize different unit processes. For example, the third largest process of ReCiPe, "petroleum//[RoW] petroleum and gas production, on-shore", does not appear in the prioritization list of the other LCIA indicators. That is, the most important unit processes according to different LCIA indicators will partly correlate and partly diverge.

Bearing this in mind, we have to focus on many different LCIA indicators in order to identify a robust set of unit processes.

[45] With inventory support we refer to the amount of unit processes which include an elementary flow addressing one of the characterization factors contained in a particular LCIA indicator (see Table S4 in the SI).

3.3.2 Prioritizing across the full set of LCIA indicators

3.3.2.1 Unit processes in the prioritization space

We use the algorithm elaborated in section 2.2 in order to identify, for each LCIA indicator, the *minimum* amount of processes *required to exceed* a given threshold. Such processes are passed into the prioritization space. Next, we count the distinct set of processes in the prioritization space (net) and their LCIA support which indicates how many LCIA indicator point to a distinct process. Figure 3-4 shows the development of the net amount of processes in the prioritization space and the corresponding LCIA support as a function of the threshold for both causer and connector perspective.

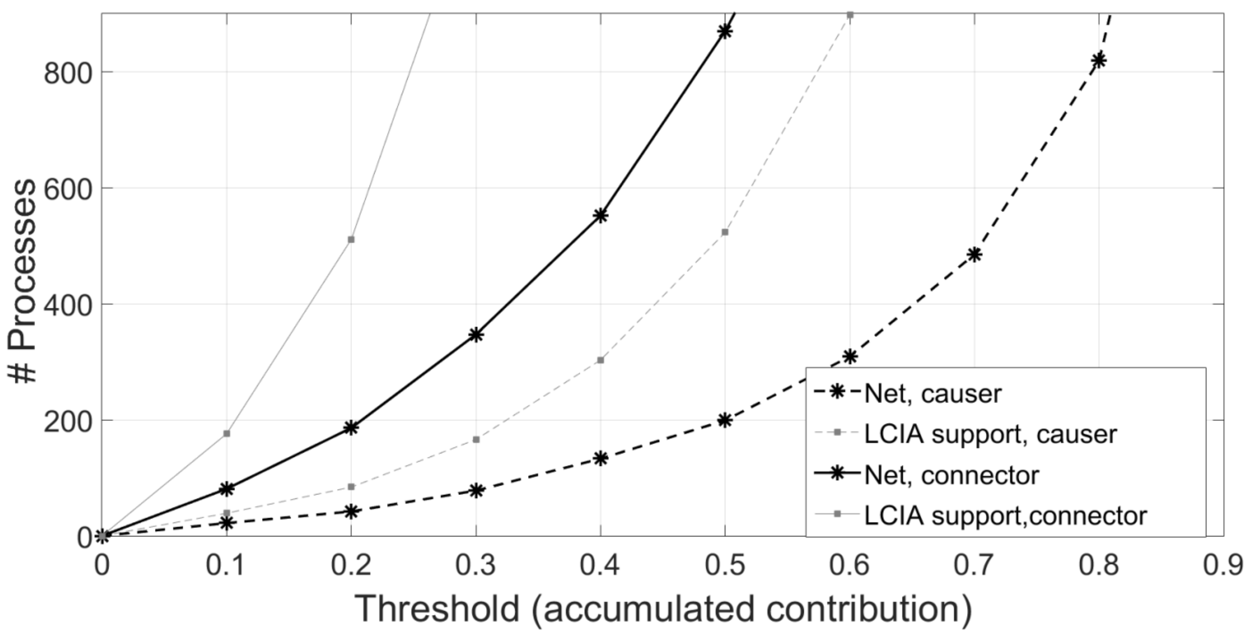


Figure 3-4: Amount of processes as a function of the threshold considered across the full set of LCIA indicators for the causer and connector perspective. Reading example for the causer perspective: to exceed a cumulated contribution of 60% across all LCIA indicators 898 unit processes are required (dashed line with squares). Since 589 are duplicates, i.e., are repeatedly added to the prioritization space by different LCIA indicators, only 309 unit processes have to be focused effectively. In other words, a tiny fraction of 309 unit processes (net, dashed line with asterisk) has an LCIA support of 898.

Figure 3-4 shows a rapid increase in the amount of duplicates in the prioritization space (i.e., the gap between the net and the LCIA support) with increasing size of the threshold. Even across many LCIA indicators the same unit processes are consistently important. These duplicates decrease the potential improvement efforts remarkably. By focusing on roughly 300 causers, we can review more than 60% of the cumulated contribution across all of the applied LCIA indicators. The connector’s perspective requires roughly an order of magnitude more processes to reach a certain threshold. As mentioned prior, almost every process in the database includes intermediate inputs that transmit environmental impacts (see Table S3-6 in section 3.7.2.1). By focusing on roughly 870 of the largest connectors we can cover 50% of the cumulated contribution across all applied LCIA indicators.

Table 3-3 shows, for each process in the prioritization space, the LCIA support. The table represents an extract for the causer perspective. The full table and the table for the connector perspective both are shown in the SI2. Table 3-3 is sorted stepwise; i.e., first according to the LCIA support in the first threshold, then according to the LCIA support in the second threshold and so forth, until the last threshold. The sorting procedure ensures that the order of processes resembles their actual relevance according to their different size classes across all LCIA indicators. We show the overall LCIA support for each process across the full set of LCIA indicators.

According to this aggregation procedure, the unit process “electricity, high voltage//[CN] electricity production, hard coal” is the most important unit process in the database as it has a LCIA support of 10, meaning that 10 LCIA indictors point to this process to exceed the 10% threshold. The order of the 3 following processes is not determined by the 10% but by the LCIA support in the following thresholds. Note, that the processes prioritized by just one LCIA indicator in the 10% threshold are often prioritized by another LCIA indicator in one of the higher threshold. It is also noteworthy that part of the LCIA support of each process is beyond the threshold of 0.6.

Table 3-3: The most important unit processes (causer perspective) according to our set of LCIA indicators. The results are sorted incrementally according to their LCIA support in the threshold. The full list is shown in SI2. Used geographical shortcuts: CN = China; RoW= Rest-of-the-World; GLO = Global; RME=Middle East; RAS=Asia; CA-ON=Ontario; BR=Brasil; IN=India.

Rank	Name [product //[geographical location] activity]	Total LCIA support	LCIA support per Threshold					
			0.1	0.2	0.3	0.4	0.5	0.6
1	electricity, high voltage//[CN] electricity production, hard coal	14	10	0	0	0	0	0
2	hard coal//[CN] hard coal mine operation	16	3	3	1	0	1	2
3	petroleum//[RoW] petroleum and gas production, on-shore	10	3	1	0	1	0	0
4	sulfidic tailing, off-site//[GLO] treatment of sulfidic tailing, off-site	8	3	0	0	2	1	0
5	diesel, burned in building machine//[GLO] diesel, burned in building machine	12	2	5	1	0	0	0
6	clinker//[RoW] clinker production	13	2	2	1	1	2	1
7	heat, district or industrial, other than natural gas//[RoW] heat production, at hard coal industrial furnace 1-10MW	13	1	4	3	1	0	0
8	petroleum//[RME] petroleum production, onshore	12	1	3	0	2	0	0
9	blasting//[RoW] blasting	9	1	2	1	2	0	0
10	natural gas, high pressure//[RoW] natural gas production	14	1	1	2	1	1	1
11	spoil from hard coal mining//[GLO] treatment of spoil from hard coal mining, in surface landfill	7	1	0	3	1	1	0
12	slag, unalloyed electric arc furnace steel//[RoW] treatment of slag, unalloyed electric arc furnace steel, residual material landfill	8	1	0	1	0	0	0
13	copper//[RAS] copper production, primary	12	1	0	0	1	2	2
14	copper//[RoW] copper production, primary	12	1	0	0	1	1	2
15	sawlog and veneer log, softwood, measured as solid wood under bark//[RoW] softwood forestry, pine, sustainable forest management	4	1	0	0	1	1	1
16	electricity, high voltage//[CA-ON] electricity production, nuclear, pressure water reactor, heavy water moderated	6	1	0	0	0	1	1
17	tailing, from uranium milling//[GLO] treatment of tailing, from uranium milling	5	1	0	0	0	0	2
18	soybean//[BR] soybean production	14	1	0	0	0	0	0
19	high level radioactive waste for final repository//[RoW] treatment of high level radioactive waste for final repository	1	1	0	0	0	0	0
20	water, decarbonised, at user//[RoW] water production and supply, decarbonised	2	1	0	0	0	0	0
21	zinc concentrate//[GLO] zinc-lead mine operation	13	1	0	0	0	0	0
22	oxygen, liquid//[RoW] air separation, cryogenic	2	1	0	0	0	0	0
23	electricity, high voltage//[IN] electricity production, hard coal	15	0	2	2	4	2	0
24	transport, freight, sea, transoceanic ship//[GLO] transport, freight, sea, transoceanic ship	12	0	2	2	3	1	0
25	pig iron//[GLO] pig iron production	13	0	2	1	1	0	0

As this format is difficult to interpret, we aggregate the LCIA support of these processes to their corresponding ISIC sector and geographical location.

3.3.2.2 The most important sectors

Figure 3-5 shows the proportion of LCIA support per threshold differentiated into ISIC rev.4[46] sectors, the activity-based classification system used by the ecoinvent database. For example, the first stacked bar (threshold 0.1) shows the cumulated LCIA support (39) of the (22 largest) processes (shown in Table 3-3) per ISIC sector. Each threshold illustrates only the additional LCIA support, i.e. it does not include the support of the prior threshold.

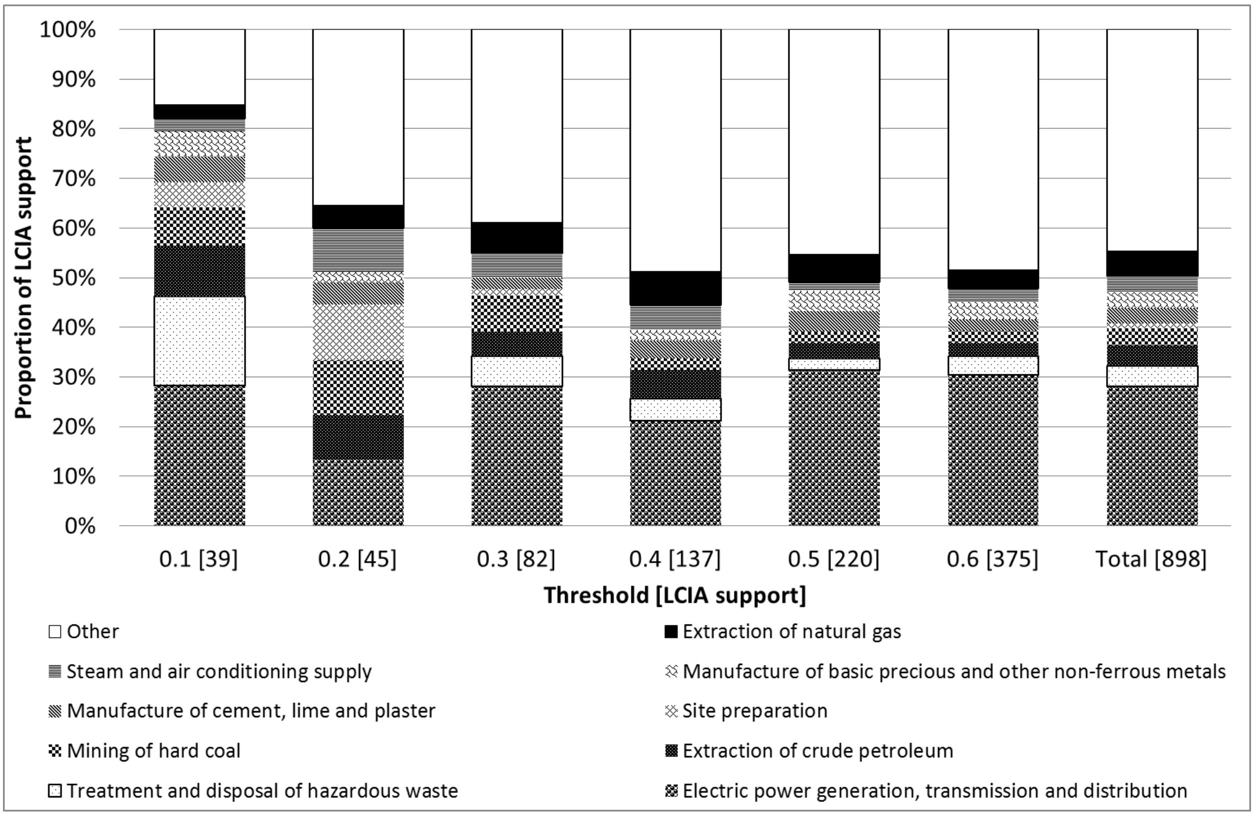


Figure 3-5: Accumulated LCIA support per threshold and ISIC rev. class. The total LCIA support per threshold is given in square brackets. For clarity, only 9 ISIC categories are shown. The remaining 35 categories are consolidated into the "Other„ category.

Figure 3-5 shows that processes referring to electric power generation, treatment and disposal of hazardous waste, extraction of crude petroleum and mining of hard coal have the largest overall effect on database quality. Overall, electric power generation causes roughly 30% of the total LCIA support. The connector perspective also highlights the importance of the electricity sector for the transmission of environmental impacts (see Figure S3-14 in section 3.7.2.6). Processes related to electric power

[46] International Standard Industrial Classification of All Economic Activities, Revision 4 (see http://unstats.un.org/unsd/cr/registry/isic-4.asp)

generation, manufacture of basic iron and steel, treatment and disposal of hazardous waste, extraction of crude petroleum, mining of hard coal and freight transport are of highest importance. They accumulate roughly 70% of the total LCIA support in the first threshold. Surprisingly, market datasets constantly represent only around 40% of the connectors throughout the first thresholds (see SI2).

Figure 3-5 also visualizes the sector relevance throughout the thresholds. From the total LCIA support added throughout the thresholds, between 12% (second threshold) and 32% (fifth threshold) refer to "electric power generation…" indicating its consistent relevance across all thresholds. The same insight applies to the extraction of crude petroleum and the mining of hard coal. The treatment and disposal of hazardous waste, in turn, is particular important with regard to the first threshold. This result from the fact that the most important processes in this sector hold a large amount of contribution such like the mentioned "tailing, from uranium milling".

3.3.2.3 The most important locations

Figure 3-6 shows the proportion of LCIA support per threshold differentiated into geographical locations.

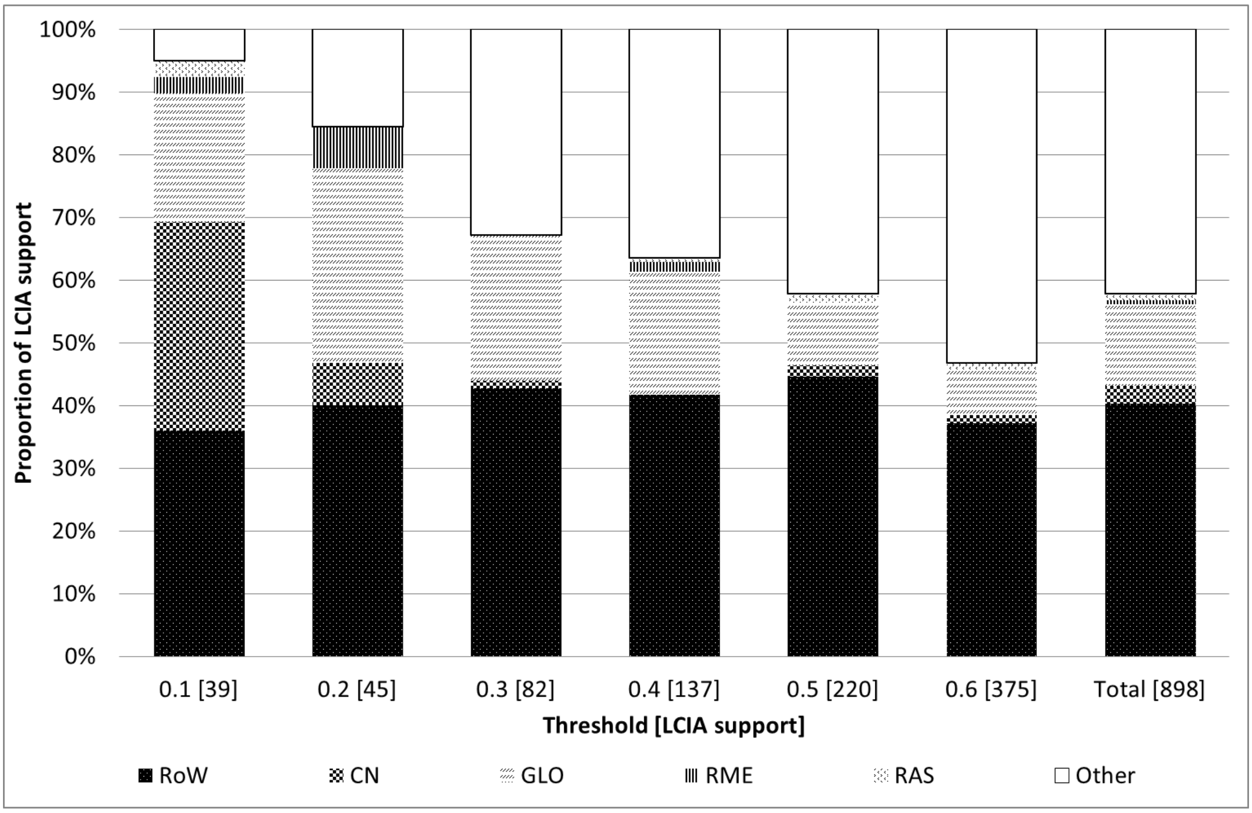


Figure 3-6: Accumulated LCIA support per threshold and geographical location. The total LCIA support per threshold is given in square brackets. Abbreviations: RoW= Rest-of-the-World; CN = China; GLO = Global; RME=Europe; RAS=Russia. The remaining 44 geographical locations are consolidated under "Other".

Figure 3-6 shows that the most important causers predominantly relate to geographical locations with a low spatial specificity, i.e. RoW, GLO and CN. These locations are particularly relevant for the first threshold where they represent 90% of the LCIA support, i.e., 36%, 33%, 21% for RoW, CN and GLO, respectively. The dominance of CN and GLO geographies decreases with increasing size of the threshold whereas the RoW is consistently important across all thresholds. The reason for this is that in many cases local production datasets only cover a small proportion of the global production volumes. The importance of RoW datasets therefore indicates a general lack of regionally appropriate data.

The connector perspective is also dominated by processes referring to RoW, GLO and CN (see Figure S3-15 in section 3.7.2.7). Processes referring to these geographies cause almost 90% of the LCIA support in the first threshold, i.e., 27%, 22%, 40% for RoW, CN and GLO, respectively. Overall, datasets belonging to these geographies generate about 62% of the total LCIA support.

3.4 Discussion

The application of our prioritization approach to the ecoinvent database shows that a tiny fraction of unit processes shows contribution patterns that are consistently large, even across many LCIA indicators. That is, a relatively large proportion of the overall database quality is dependent on a small set of processes. Concentrating research efforts on the increase of information density in these systemic processes offers important starting points for the systematic and effective improvement of the entire database.

We investigated two prioritization perspectives which support the detection of two different characteristics one and the same unit process can have. The causer perspective highlights processes with exchanges of resources and emissions that are consistently important across product system and LCIA indicators. Overall, 3% of the processes in the database cause more than 60% of the total mean contribution. Processes referring to electricity generation, waste treatment activities and energy carrier provision (petroleum and hard coal) are consistently important. The connector perspective emphasizes sensitive hubs whose modification can alter the results for the overall database considerably. In total, 8% of the processes transmit more than 50%. It is worth recalling, however, that these 50% are normalized to the contribution transmitted by all connectors. Effectively, the contribution "flowing through" these most important connectors translates into roughly 4 times the total contribution in the database. This highlights the overall importance of such hubs for database quality and suggests that they receive more attention. Particularly, electricity generation but also iron and steel production has strong networking effects.

The relatively low amount of systemic datasets should not be automatically taken as a general characteristic of LCA or the economic system it strives to model. In fact, extra to the already mentioned large correlation among LCIA indicators, it is also the build-up of partly arbitrary and partly intended modeling choices that incorporates a certain level of inequality into the database structure.

- First, it is the presence of different process types that fosters a certain degree of inequality in process importance. In addition to transforming processes, ecoinvent has always used consolidating processes to represent certain geographical and technological averages or to maintain a certain modeling structure. Typically, such datasets cause no direct environmental intervention and consequently have no contribution in the causer perspective. Version 3.1 of the database maintains 3'196 market datasets. That is, a large proportion of the processes in the database are simply not relevant for the causer perspective (see Table S3-6 in section 3.7.2.1).
- Second, some elementary flows (resources or emissions) are only listed in a few unit processes and consequently the inventory support for some LCIA indicators is rather low (see Table S3-7 in section 3.7.3.1). This is not necessarily a characteristic of our physical reality but results, at least in

part, from the fact that more attention was paid towards the cultivation of some emissions or impact categories than others. For example, among the assessed mid-point methods, LCIA indicators related to climate change have the largest inventory support – 3’500 processes cause emission with a global warming potential – while agricultural land occupation and resource depletion indicators appear at the other end of the scale with an inventory support of roughly 300 processes. The endpoint method with the largest inventory support is the Swiss ecological scarcity method (2013) with roughly 6’000 processes. The inventory support seems to be a good predictor for the size differences in the overall inequality among LCIA indicators, i.e. the degree of concentration in the mean process contribution vectors. The correlation is high for both the causer and the connector perspective (see the correlation analysis in Table S3-8 in section 3.7.3.1). Note that the latter (plus the high correlation between the causer and the connector perspective) suggest that the inequality in the causer perspective is, more or less directly, passed to the connector perspective.

- A third reason for the large inequality is related to the fact that the ecoinvent database has evolved as (and still is) an incomplete model of our economic system. The incomplete spatial and technological process support has at least two notable effects. First, it limits the selection of appropriate intermediate inputs required for the accurate representation of new processes. This fosters, at least to a certain degree, the iterative use of the same (generic) intermediate inputs and favors a network structure with consistent dependencies. For example, the importance of the process “diesel, burned in building machine//[GLO]” (rank 5) results, at least in part, from its lack in spatial and technological specificity and the lack of alternatives. Second, the incompleteness in spatial process support causes (since version 3.0) the repetitive linking to generic geographical locations and consequently promotes the importance of such locations. Whenever local datasets are unavailable for the supply of a certain product, the product is supplied by the GLO dataset. If local production datasets are available but only cover a small proportion of the global production volumes, most of the production volume is supplied by a RoW dataset. That is, the overrepresentation of generic geographical locations in the first threshold indicates a clear lack in spatial process support.

We excluded all non-contributing processes from the mean contribution vectors in order to get a first idea of the influence reason 1 and 2 have on the overall inequality. The inequality in the mean contribution vectors decreases but remains high (see Table S3-7 in section 3.7.3.1). Furthermore, the exclusion of non-contributing processes cancels the correlation between the inventory support and the inequality (see Table S3-8 in section 3.7.3.1). This suggests that (i) the inventory support (or lack therefore) indeed drives differences among the inequality in the LCIA indicators and (ii) the remaining, large inequality in process importance relates mainly to the incompleteness of the database (reason 3) and other, non-identified reasons[47]. Both indicate the significance of advancements in the spatial process support (specificity) of the database but also in the further refinement of process completeness.

Some issues of importance need to be considered with regard to the prioritization method. First, our approach represents an internal perspective where the optimal allocation follows the structural importance of the unit processes according to their relative mean contribution across different LCIA indicators. This is useful for the identification of consistently important processes but provides no direct support for the identification of "blank spots" and is therefore "unable to direct research efforts to economic sectors that may be underrepresented" (Reinhard et al., 2016). Therefore, its inward-oriented perspective should be complemented with more outward-oriented prioritization methods such like the one presented by (Majeau-Bettez et al., 2011). Note, however, that the method offers some indirect support for the identification of blank spots, namely via the identification of processes which are modelled to generic and which should be modelled more specifically in spatial or in technological scale. A typical example is the already mentioned process "diesel, burned in building machine//[GLO]".

Further, our consolidation algorithm ensures for a given set of LCIA indicators that all processes required to exceed a certain threshold are considered. It operators on the basis of the Lorenz curve which is the most prominent tool for analyzing and comparing income inequality (Duclos and Araar, 2006). However, the discretization to size classes comes at the cost of information loss, notably large contributions are reduced to the size class of the threshold. In general, as the order of processes is first determined by their actual size class (threshold) and only then by their LCIA support, the sorting procedure maintains essential information about the actual size of a unit process contribution. It ensures that processes in the first threshold, only prioritized by one rather uncorrelated LCIA indicator with a low LCIA support (such like ALOP or IR), will still receive more attention than a process with a high LCIA support in the second threshold. We recognize the potential utility of this consolidation procedure within every LCA application as the avoidance of "burden shifting", inherent to every LCA, often involves the identification of the most important processes according to many LCIA indicators (Hellweg and Milà i Canals, 2014). That is, the consolidation algorithm could support the interpretation phase of every standard LCA. The selection of the thresholds in this article is rather coarse in order to maintain easy interpretation. In order to achieve a more accurate ranking of processes, the complete list of ranked processes presented in SI2, uses a much finer threshold scale of 0.01. Note, however, that this only affects the ranking of the

[47] In general, it is difficult to separate the actual influence of (i) the inequality embedded in our economic system and (ii) the incompleteness of the database. We believe that a large proportion of the reaming inequality is caused by the incompleteness of the database because of the mentioned overrepresentation of generic geographical locations and technologies.

processes within a particular threshold but not the actual amount of processes associated with this threshold. That is, the same 22 unit processes shown in Table 3-3 are required to exceed the threshold of 10%.

The discretization also guarantees a basic equivalence among the LCIA indicators. While this avoids (implicit) weighting of LCIA indicators on the basis of the inequality in their mean contribution vector, it means that each LCIA indicator is treated equally. That is, a unit process added by the midpoint indicator climate change (IPCC2007) is considered of equal importance than a unit process added by the endpoint indicator ReCiPe total. One can imagine multi-criteria approaches which assign different weights to different LCIA indicators. Note, however that, as we keep all unique unit processes on the prioritization list, such a weighting would only affect the rank of the processes in the list but not change the actual amount of processes on the list.

The actual amount of processes on the list is dependent on the set of selected LCIA indicators. Each LCIA indicators can be considered as a unique optimization vector. Therefore, adding or removing LCIA indicators from the selected set can change the amount of prioritized processes. Given the premise of random LCIA indicators selection, Reinhard (2016) shows that the average dependence on LCIA indicators diminishes with the total amount of LCIA indicators considered. That is, due to the generally high correlation among LCIA indicators the marginal addition of unit processes (to the already existing prioritization list) caused by e.g. the consideration of a 20th LCIA indicator is much lower than the addition of a 5th. Consequently, adding five randomly selected LCIA indicators to the existing set of LCIA indicators will, in general, not add many processes to the list as the combined inventory support of the set will not change significantly. In general, the prioritization method is applicable to any set of LCIA indicators and future work should carefully analyze different LCIA indicators perspectives, e.g. a set of midpoint methods versus a set of endpoint methods, and the corresponding differences in their recommendations. This would reveal the detailed inventory support of different LCIA paradigm and therefore offer important feedback for LCIA method developers.

Finally, the present approach does not consider the uncertainty dimensions. Computing the mean contribution to uncertainty for each process throughout all product systems and contrasting it with the results from the presented contribution analysis would support a much more fine-grained configuration of improvements and therefore further advance the effectiveness of the method. The computational costs of such an analysis are expected to be very large as it could require several ten thousands of matrix inversions and conversions of the technosphere matrix which, for ecoinvent version 3.2., has a size of roughly 13,000. The topic was beyond the scope of this article but deserves future investigation.

3.5 Conclusion

Improving the quality and quantity of unit process datasets in LCI databases affects every LCA they are used in. Precise knowledge of the processes with the largest contribution is one prerequisite of systematic, targeted improvement. We demonstrated how a contribution-based approach can be applied using version 3.1 of the ecoinvent database. The approach facilitates the prioritization of areas for improvement under the joint perspective of a broad range of LCIA indicators. The presented list of most important processes offers new starting points for the effective improvement of the ecoinvent database as it allows the allocation of available resources according to the structural dependencies in the data. This strengthens the basis for decision making of the people managing the data, who can systematically process the list depending on the available time and resources and decide for each unit process if and how it should be improved. The high inequality in process importance offers a new starting point for the definition of minimum quality requirements for systemic unit process datasets.

The efficient allocation of available resources is a key issue for the improvement of LCI databases. We observe a general lack of operational tools able to express important characteristics of an entire LCI database and to realign the focus of inventory efforts to more systematic instead of arbitrary choices guided by external data availability and situation-driven requirements. Certainly, such data and requirements have a great importance, but future research should also ensure that coverage and specificity of LCI databases progress more effectively towards a model of the complex economic system LCA strives to capture. The structural dependencies in the database, as discussed in this paper, represent a meaningful improvement perspective but should be complemented with other, more outward-looking perspectives.

3.6 References

Bourgault G, Lesage P, Samson R (2012) Systematic disaggregation: a hybrid LCI computation algorithm enhancing interpretation phase in LCA. Int J Life Cycle Assess 17:774–786. doi: 10.1007/s11367-012-0418-7

Ciroth A, Muller S, Weidema B, Lesage P (2013) Empirically based uncertainty factors for the pedigree matrix in ecoinvent. Int J Life Cycle Assess 1–11. doi: 10.1007/s11367-013-0670-5

Duclos J-Y, Araar A (2006) Poverty and Equity: Measurement, Policy, and Estimation with DAD. Springer

Ebner F (2013) Results, Interpretation and Reporting: GaBi Software. In: Gabi-Softw. http://www.gabi-software.com/software/gabi-software/gabi-5/functionalities/results-interpretation-and-reporting/. Accessed 16 Aug 2013

EC-JRC (2011) International Reference Life Cycle Data System (ILCD) Handbook- Recommendations for Life Cycle Impact Assessment in the European context. First edition November 2011. EUR 24571 EN. Luxemburg. Publications Office of the European Union;

Finnveden G, Hauschild MZ, Ekvall T, et al (2009) Recent developments in Life Cycle Assessment. J Environ Manage 91:1–21. doi: 10.1016/j.jenvman.2009.06.018

Goedkoop M, Oele M (2004) SimaPro 6 - Introduction to LCA with SimaPro.

Heijungs R, Huijbregts M (2004) A review of approaches to treat uncertainty in LCA.

Heijungs R, Kleijn R (2001) Numerical approaches towards life cycle interpretation five examples. Int J Life Cycle Assess 6:141–148. doi: 10.1007/BF02978732

Hellweg S, Milà i Canals L (2014) Emerging approaches, challenges and opportunities in life cycle assessment. Science 344:1109–1113.

Huijbregts MJ (1998) Application of uncertainty and variability in LCA. Int J Life Cycle Assess 3:273–280. doi: 10.1007/BF02979835

Majeau-Bettez G, Strømman AH, Hertwich EG (2011) Evaluation of Process- and Input–Output-based Life Cycle Inventory Data with Regard to Truncation and Aggregation Issues. Environ Sci Technol 45:10170–10177. doi: 10.1021/es201308x

Reinhard J, Mutel C, Wernet G, et al (2016) Contribution-based prioritization of LCI database improvements: Method design, demonstration, and evaluation. Environ Model Softw 86:204–218. doi: 10.1016/j.envsoft.2016.09.018

Sonnemann G, Vigon B (eds). (2011) Global Guidance Principles for Life Cycle Assessment Databases: A Basis for Greener Processes and Products (Shonan Guidance Principles).

Steubing B, Wernet G, Reinhard J, et al (2016) The ecoinvent database version 3 (part II): analyzing LCA results and comparison to version 2. Int J Life Cycle Assess 21:1269–1281. doi: 10.1007/s11367-016-1109-6

Tillman A-M (2000) Significance of decision-making for LCA methodology. Environ Impact Assess Rev 20:113–123. doi: 10.1016/S0195-9255(99)00035-9

Wernet G, Bauer C, Steubing B, et al (2016) The ecoinvent database version 3 (part I): overview and methodology. Int J Life Cycle Assess 21:1218–1230. doi: 10.1007/s11367-016-1087-8

3.7 Supplementary Information (II 1)

3.7.1 Methods

3.7.1.1 Variables and computations to calculate relative contribution matrixes

Table S3-4: Matrixes and vectors required to calculate relative contribution matrixes. Source: (Reinhard et al., 2016).

No.	Name	Symbol	Content	Dimension	Source/Computation
1	Technosphere matrix	A	Intermediate product consumptions by processes.	$n \times n$	Supplied by the database
2	Biosphere matrix	**B**	Elementary exchanges (emissions and natural resources) by processes.	$i \times n$	Supplied by the database
3	Identity matrix	**I**	Ones on the main diagonal and zeros elsewhere. Required to establish (explicitly) a unitary demand for each product.	$n \times n$	Set up by LCA practitioner
4	Characterization matrix	**W**	Characterization factors by elementary exchanges.	$k \times i$	Supplied by the database
5	Supply (or scaling)	**S**	Process amount (required to satisfy a unitary	$n \times n$	$\mathbf{S} = \mathbf{A}^{-1}\mathbf{I}$

	matrix		product demand).by products.		
6	Characterization vector	w	Characterization vector from the characterization matrix $\mathbf{W}$, i.e. one (selected) LCIA indicator.	$1 \times i$	$w = \boldsymbol{W_i}$
7	Direct contribution vector	e	Direct environmental impacts by processes. Represents for each process the environmental impacts caused directly by its elementary flows.	$1 \times n$	$e = w\mathbf{B}$
8	Direct contribution matrix	**Md**	Direct, process-specific environmental impacts by products. Shows the direct impact of each process throughout all product systems.	$n \times n$	$\mathbf{Md} = \hat{e}\mathbf{S}$
9	Total impact vector	td	Total environmental impact by products. Shows the total impacts over the entire product system, associated with a unitary product demand.	$1 \times n$	$\mathrm{td}_j = \sum_i \mathbf{Md}_{ij}$
10	Expansion vector	$\vec{1}$	A vector of ones used to expand vectors td and p to matrix form.	$n \times 1$	Set up by LCA practitioner
11	Relative direct contribution matrix	**Rd**	Relative, process-specific contributions by products. Shows the relative contribution of each process throughout all product systems.	$n \times n$	$\mathbf{Rd} = \mathbf{Md} \oslash (\vec{1}td)$
12	Cumulated impact per process	tn	Cumulated (direct and upstream) environmental impact by processes. Shows the impacts associated with the provision of one unit of each process.	$n \times 1$	$tn = td^{\mathrm{T}} \oslash \hat{\mathbf{S}}$
13	Cumulated contribution matrix	**Mc**	Cumulated (direct and upstream) environmental impact of processes by products.	$n \times n$	$\mathbf{Mc} = \widehat{tn}\mathbf{S}$
14	Upstream proportion vector	p	Proportion of process-specific environmental impacts caused upstream.	$1 \times n$	$p = 1 - (e \oslash td)$
15	Upstream contribution matrix	**Mup**	Upstream environmental impact of processes by products. Shows the upstream contribution of each process throughout all product systems.	$n \times n$	$\mathbf{Mup} = \mathbf{Mc} \circ (\vec{1}p)^{\mathrm{T}}$
16	Relative upstream contribution matrix	**Rup**	Relative, process-specific upstream contribution by products. Shows the relative upstream contribution of each process throughout all product systems.	$n \times n$	$\mathbf{Rup} = \mathbf{Mup} \oslash (\vec{1}td)$
17	Mean contribution vector	**Ø**	Arithmetic mean of all contributions of a process throughout all product system.	$n \times 1$	$\text{Ø}\mathbf{d}_i = \frac{1}{n}\sum_j \mathbf{Rd}_{\mathrm{ij}}$ $\text{Ø}\mathbf{up}_i = \frac{1}{n}\sum_j \mathbf{Rup}_{\mathrm{ij}}$
18	Normalized mean contribution vector	**ØN**	Normalized arithmetic mean	$n \times 1$	$\text{Ø}\boldsymbol{N}_i = \frac{\text{Ø}_i}{\sum_i^n \text{Ø}_i}$
19	Gini coefficient	$\boldsymbol{Gc}$	Expresses the inequality in the mean contribution vector of a particular LCIA indicator as the "relative mean difference, i.e., the mean of the differences between every possible pair of processes, divided by the mean size" (Damgaard and Weiner, 2000, p. 1139).	$k \times 1$	$\boldsymbol{Gc} = \frac{\sum_{i=1}^n \sum_{j=1}^n \lvert y_i - y_j \rvert}{2n^2\mu}$

3.7.1.2 Lorenz curve

We use the mirror image of the Lorenz curve (MLC) to express the cumulative percentage of the total contribution held by a cumulative proportion of processes. That is, instead of plotting the lowest contribution first, we start with the process whose mean contribution throughout the database is the largest and proceed by adding all other mean process contributions in descending order. More formally, if we have n ordered process contributions, such that $\emptyset_i$ is the mean of process i and $\emptyset_1 \geq \emptyset_2 \geq \cdots \geq \emptyset_n$, then the discrete MLC is defined as the polygon joining the points $\left(p = \frac{h}{n}, c = \frac{ML_h}{ML_n}\right)$ where $h = 1, 2 \dots n$, $p_0 = 0$, $c_0 = 0$, $ML_n = \sum_{i=1}^{n} \emptyset_i$ and $ML_h = \sum_{i=1}^{h} \emptyset_i$ (Damgaard and Weiner, 2000).

3.7.1.3 Our set of LCIA indicators

We selected 19 LCIA indicators (Table S3-5) from seven LCIA methods based on their scientific quality[48], their availability[49], and their assumed relevance in practice.

Table S3-5: LCIA indicators applied. The LCIA methods marked with an asterisk are used for a detailed analysis (see section 3.1).

No.	LCIA method	Indicator	Unit	Abbr.
1	IPCC 2007*	Climate Change, GWP100a	kg CO2 eq	CC
2	CML 2001	Resources, depletion of abiotic resources	kg antimony eq	ARD
3	ReCiPe Midpoint (H)	Photochemical oxidant formation	kg NMVOC	POF
4	ReCiPe Midpoint (H)	Marine eutrophication	kg N eq	MEP
5	ReCiPe Midpoint (H)	Ionizing radiation	kBq U235 eq	IR
6	ReCiPe Midpoint (H)	Ozone depletion	kg CFC-11 eq	ODP
7	ReCiPe Midpoint (H)	Freshwater eutrophication	kg P eq	FEP
8	ReCiPe Midpoint (H)	Agricultural land occupation	m^2a	ALOP
9	ReCiPe Midpoint (H)	Terrestrial acidification	kg 1,4-DB eq	TAF
10	USEtox	Human toxicity, carcinogenic	CTUh	HtoxC
11	USEtox	Human toxicity, non-carcinogenic	CTUh	HtoxNC
12	USEtox*	Ecotoxicitiy, total	CTUe	Etox
13	IMPACT 2002+	Resources	MJ primary	I+R
14	IMPACT 2002+	Ecosystem quality	PDF*m2*yr	I+EQ
15	IMPACT 2002+	Human health	DALY	I+HH
16	IMPACT 2002+	Climate change	kg CO_2 eq	I+CC
17	Ecological Scarcity 2013	Total	points	ES13_total
18	Ecological Scarcity 2013	Water resources	Points	ES13Water
19	ReCiPe Endpoint (H,A)*	Total	points	ReCiPe

3.7.1.4 Visualizing the relative contribution matrixes for abiotic resource depletion (ARD)

[48] Scientific quality is determined according to the recommendations in the ILCD-Handbook (EC-JRC, 2011).

[49] Though the ecoinvent center provides the aggregated results of 692 LCIA indicators, it does not cover all LCIA indicators recommended by the ILCD-Handbook.

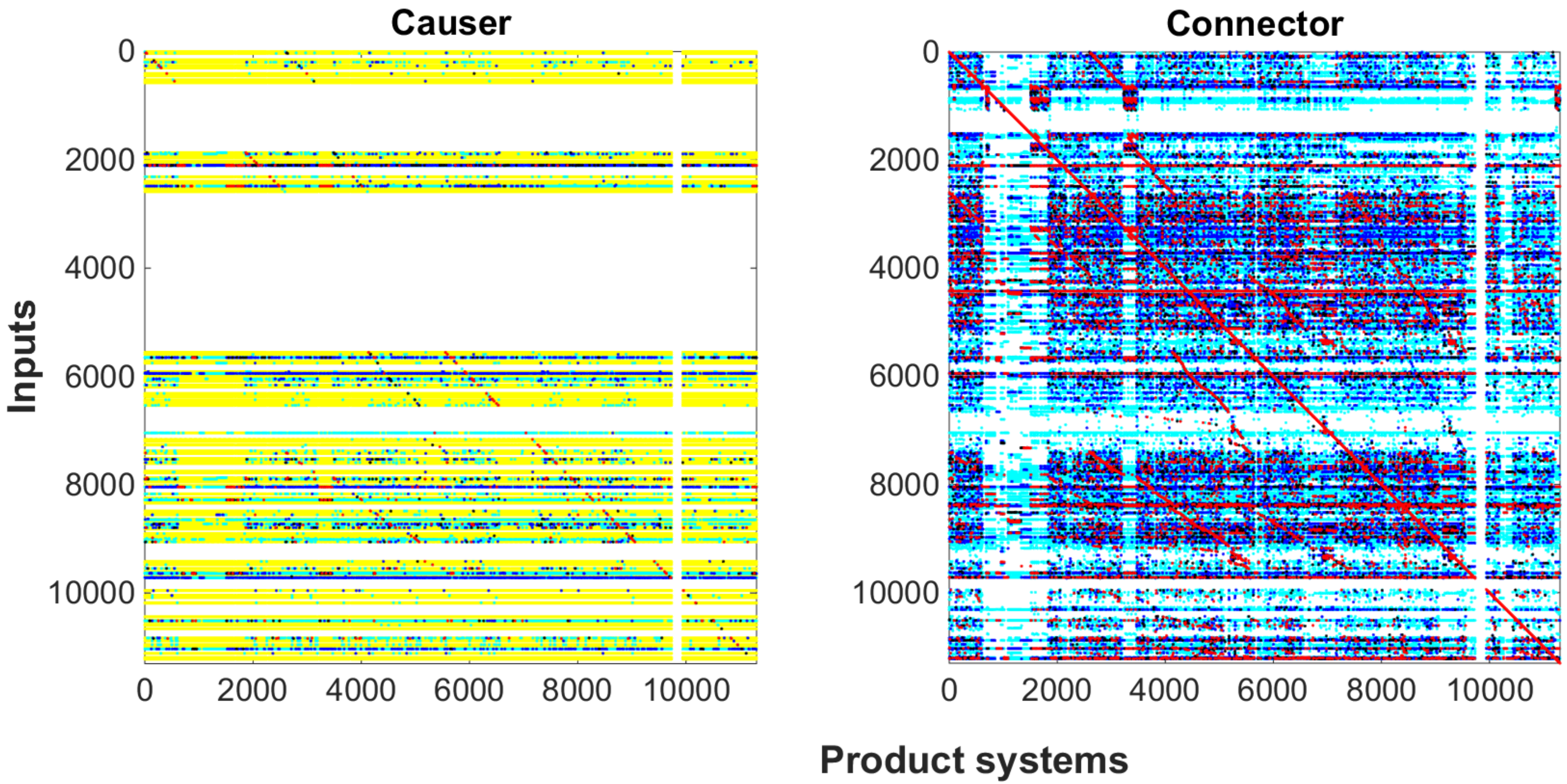


Figure S3-7: Colored-coded spy plots visualizing the use and contribution patterns of 11,304 processes throughout 11,304 product systems for the relative contribution matrixes Rd and Rup for the LCIA indicator ARD. Color code: contributions exceeding 0%, 1%, 10%, 30%, 50% are yellow, turquoise, blue, black and red, respectively. Processes with no contribution are shown as white.

3.7.2 Results

3.7.2.1 General database metrics

Table S3-6 shows some general statistics about the selected database version; Allocation, cut-off by classification" (cut-off).

Table S3-6: General database metrics of ecoinvent version 3.1, cut-off.

Database metrics	Connector	Causer	Total
# of Unit processes	11,008	6,420	11,304
Empty (cut-off) processes			296
Market to empty processes	158		158
Ø [σ] of elementary exchanges per process	13 [25]	22 [29]	12 [25]
Ø [σ] of intermediate inputs per process	23 [29]	31 [32]	23 [29]

Consistently, around 450 processes have no input from the technosphere and no elementary flows. Most of these processes (296) are empty and result from the cut-off procedure inherent to the cut-off system model. The remaining processes (158) are markets that exclusively link to these empty processes. The database contains about 11,008 connectors (11,304 - 296), i.e., unit processes that include at least one intermediate exchange, and 6,420 causers, i.e., unit processes that include at least one elementary exchange.

The average unit process has 12 elementary exchanges and 23 intermediate inputs. Since almost all unit processes in the database are connectors this metrics does not change notably if we focus on the connectors only. However, the average causer shows quite different characteristics; it has, on average, 22 elementary exchanges and 31 intermediate inputs.

3.7.2.2 Frequency of use - Causer

Figure S3-8 shows the histogram for the frequency of use (FoU) across all assessed LCIA methods for the causer perspective.

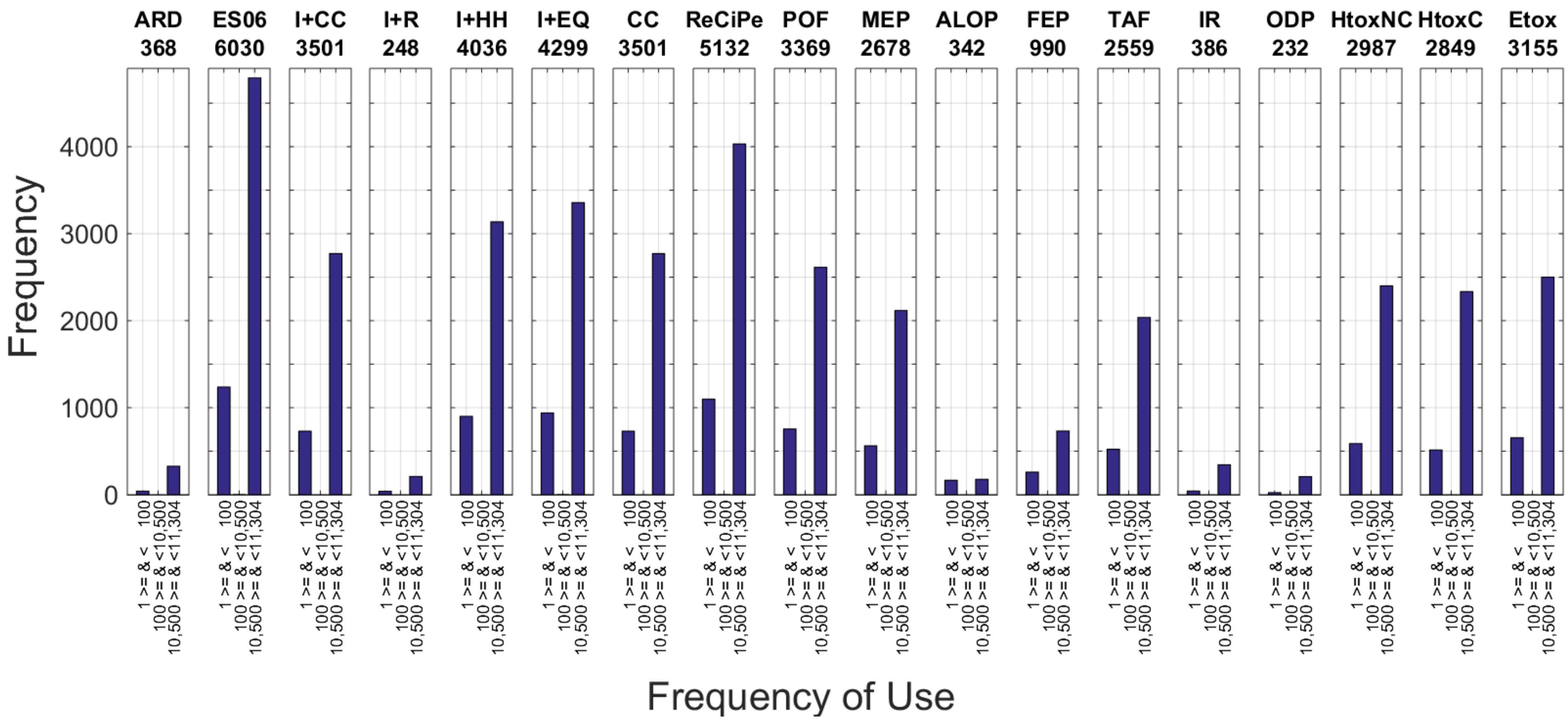


Figure S3-8: Histogram showing the frequency of processes for three bins across all LCIA methods. Reading example for ARD: 368 processes throughout the database have a mean contribution unequal to zero (n=11,304), i.e. their elementary flows are considered by ARD. 347 processes are used throughout more than 10,500 product systems, 41 processes are used throughout less than 100 product system. None of the processes shows a FoU that lies in between.

As shown by Figure S3-8, the FoU is highly dependent on the LCIA indicator. Typically more than half of the processes (which have a mean contribution unequal to zero) are highly integrated into the database meaning that they are used throughout more than 10,500 product systems, while the remaining processes show a low degree of integration, i.e. they are used throughout less than 100 product systems or not at all.

Figure S3-9 shows the FoU of each process and its corresponding mean contribution for all assessed LCIA indicators. In order to improve readability the plot is scaled in a logarithmic fashion. It shows that processes with a mean contribution exceeding 0.5% are always used by the entire database, i.e. throughout more than 10,500 product systems. Processes with a mean contribution below 0.5% are characterized by both a high and a low integration. Note, however, that a process which is used throughout 20 product systems can have a larger mean contribution than a process that is used throughout all product systems. That is, a high degree of integration (FoU) is no sufficient condition for a large mean process contribution.

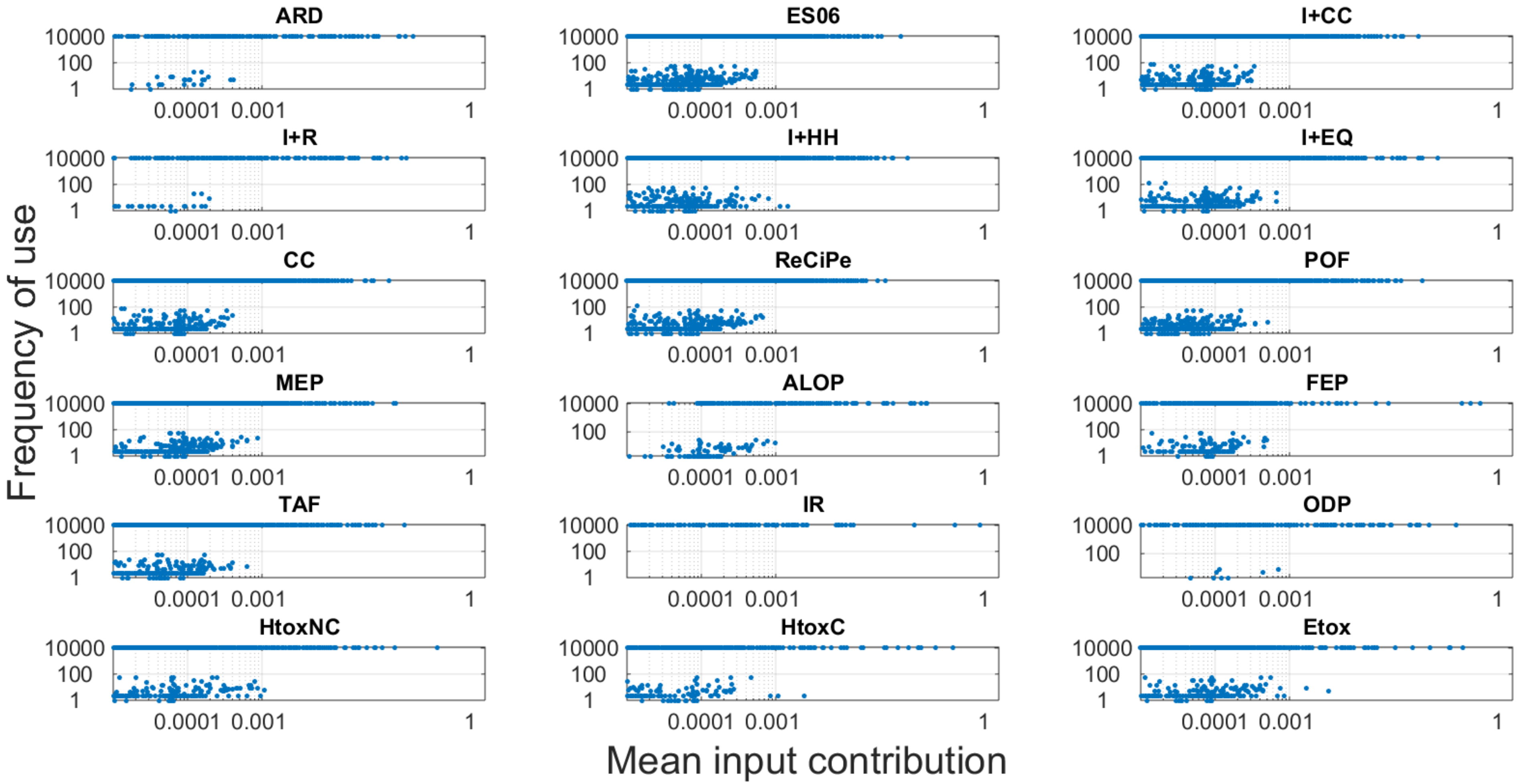


Figure S3-9: Log-log scatter plot showing the frequency of use in relation to the mean process contribution for the causer perspective. All processes with mean contribution above 0.5% are used throughout almost all product system, i.e. FoU is above 10,500.

3.7.2.3 Frequency of use - Connector

Figure S3-10 shows the histogram for the frequency of use across all assessed LCIA methods for the connector perspective. In contrast to the causer perspective, the FoU patterns of the processes in the connector perspective are equal across all LCIA methods, i.e. independent of the LCIA indicator; 40% of the processes are used very rarely (<100), while 60% of the processes are used by almost all (>10’500) product systems.

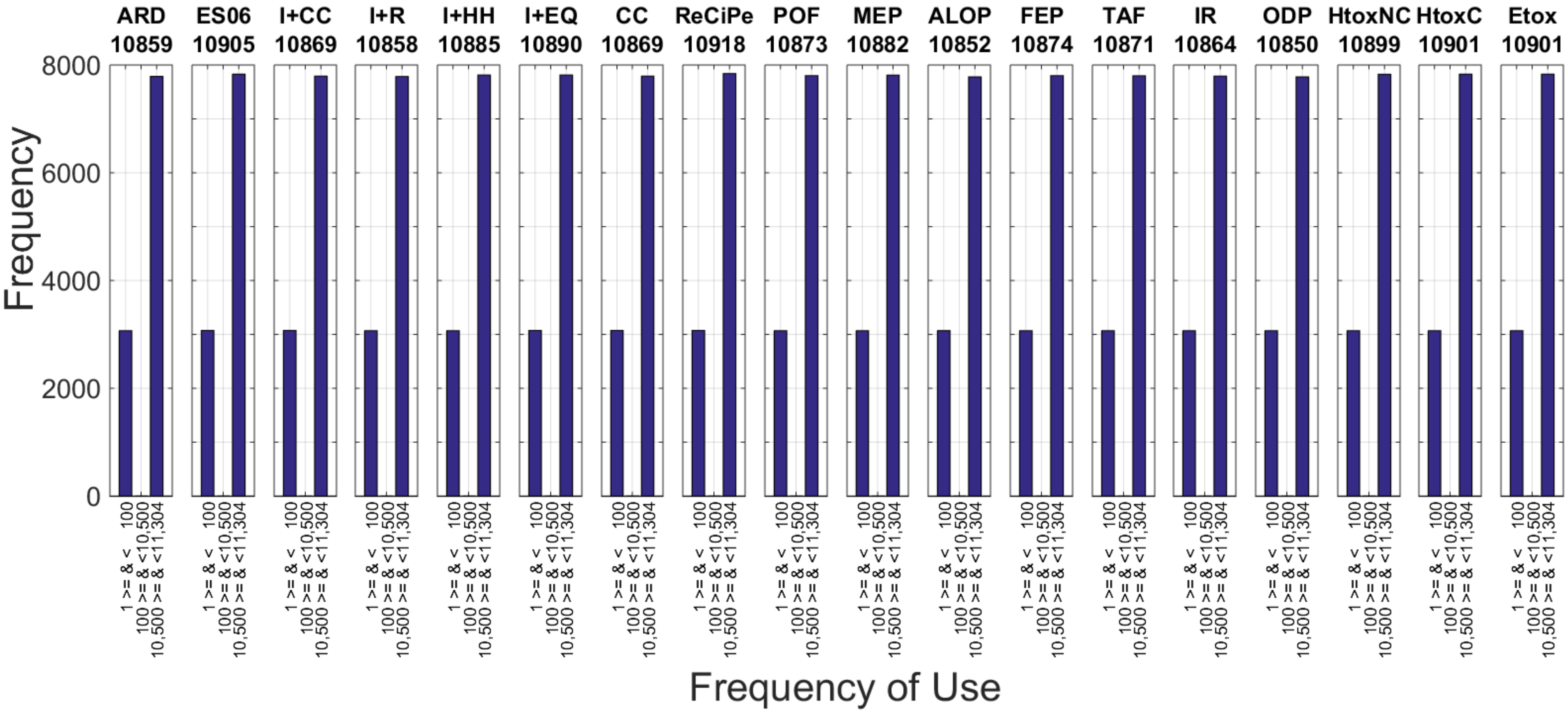


Figure S3-10: Histogram showing the frequency of processes for three bins across all LCIA methods. Reading example for ARD: 10,859 processes throughout the database have a mean contribution unequal to zero (n=11,304); 7,788 processes are used throughout more than 10,500 product systems, 3,068 processes are used throughout less than 100 product systems. None of the processes shows a FoU that lies in-between.

Figure S4 shows the FoU of each input and its corresponding mean contribution for all assessed LCIA methods. In order to improve readability the plot is scaled in a logarithmic fashion. It shows that processes with a mean contribution exceeding 0.1% are always used by the entire database, i.e. throughout more than 10’500 product systems.

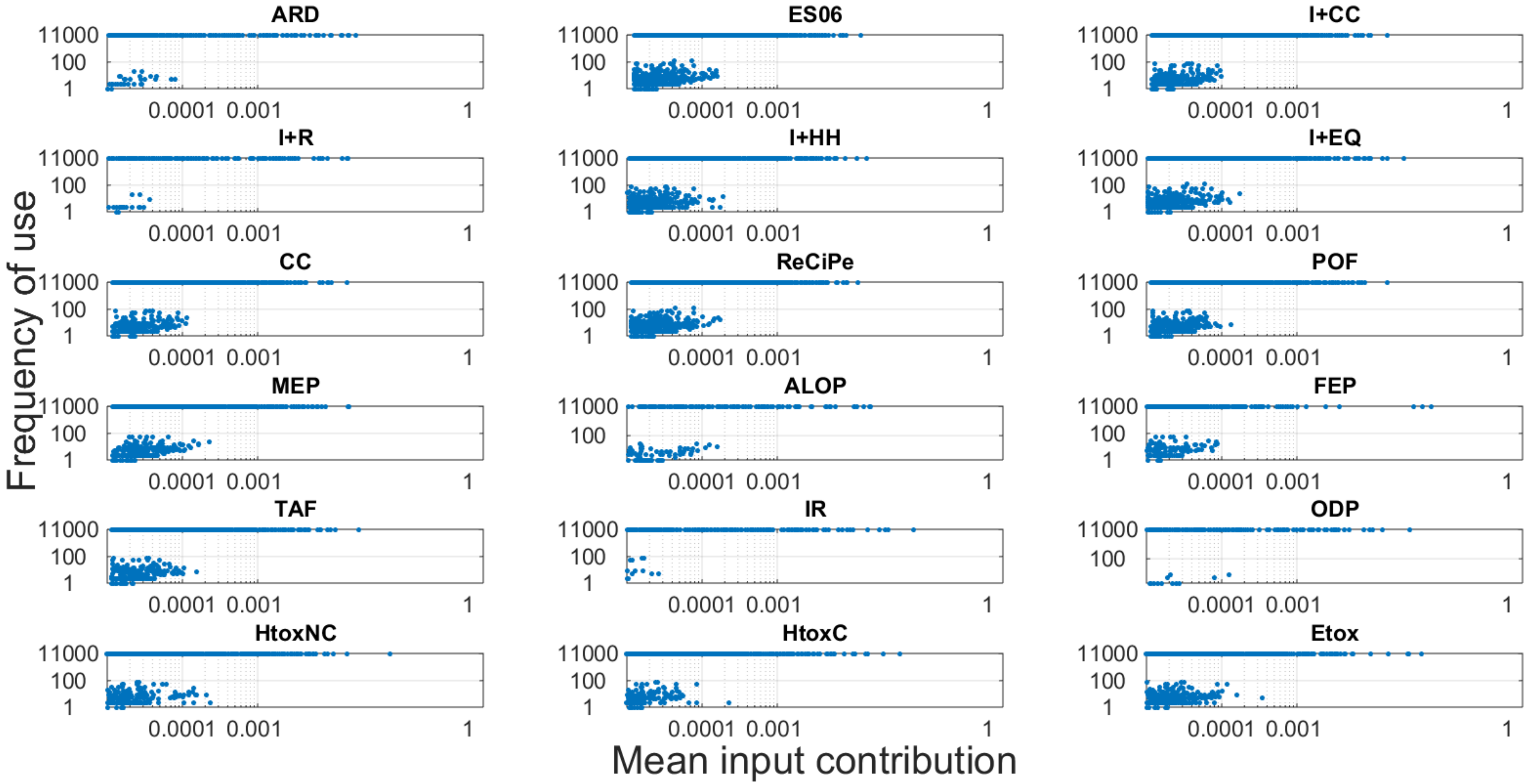


Figure S3-11: Log-log scatter plot showing the frequency of use in relation to the mean input contribution for the connector perspective. All processes with mean contribution exceeding 0.1% are used throughout almost all product system, i.e. FoU is above 10,500.

3.7.2.4 A certain degree of correlation between the causer and the connector perspective

Figure S3-12 shows that there is a certain degree of correlation between the causer and the connector perspective. At the same time, many important connectors have no contribution in the causer perspective (A)—and therefore do not appear in the logarithmic illustration. This holds true for many other connectors as well.

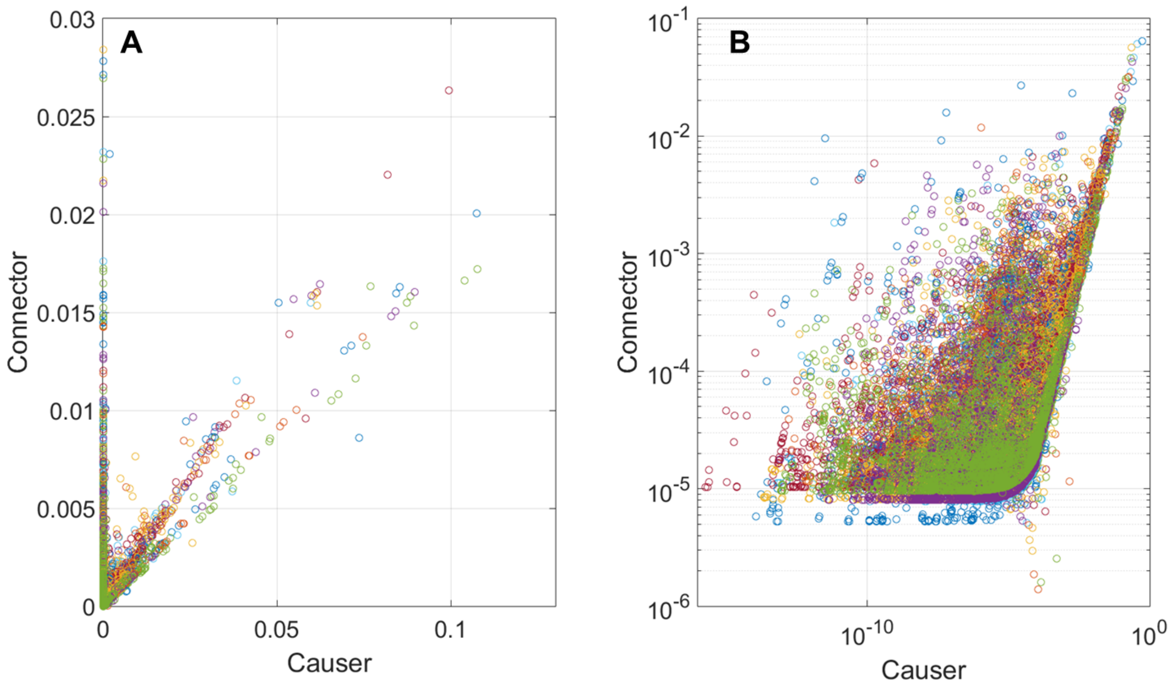


Figure S3-12: Scatter plot contrasting the mean process contributions of the causer perspective with the connector perspective for all 19 LCIA indicators with linear (A) and logarithmic (B) scales.

3.7.2.5 Connectors transmit several times the mean contribution of causers

Figure S3-13 shows the ratio between the cumulated mean contribution of the connector perspective and the cumulated mean contribution of the causer perspective for all 19 LCIA indicators. The figure is generated on the basis of the "CumulatedMeanContributionCauser" and "CumulatedMeanContributionConnec" (see SI2).

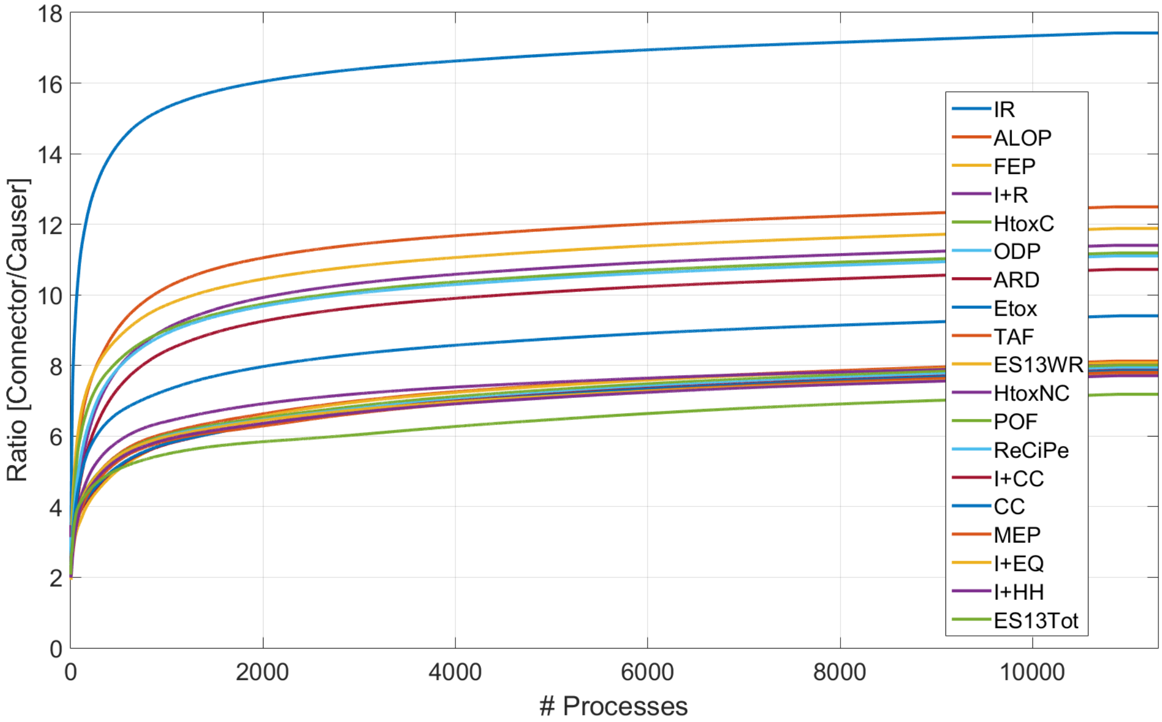


Figure S3-13: Ratio between the cumulated mean process contribution of the connectors and the causers for all analysed LCIA indictors.

As shown by Figure S3-13, depending on the LCIA indicator, the connectors transmit 7 to 17 times the cumulated contribution caused by the causers. The ratios develop in a logarithmic fashion. Typically, 1,000 out of 11,304 processes already transmit 70% of the overall contribution associated with a particular LCIA indicator.

3.7.2.6 The most important sectors

Figure S3-14 shows the proportion of LCIA support per threshold differentiated into ISIC rev.4 sectors. Each threshold illustrates only the additional LCIA support, i.e. it does not include the support of the prior threshold.

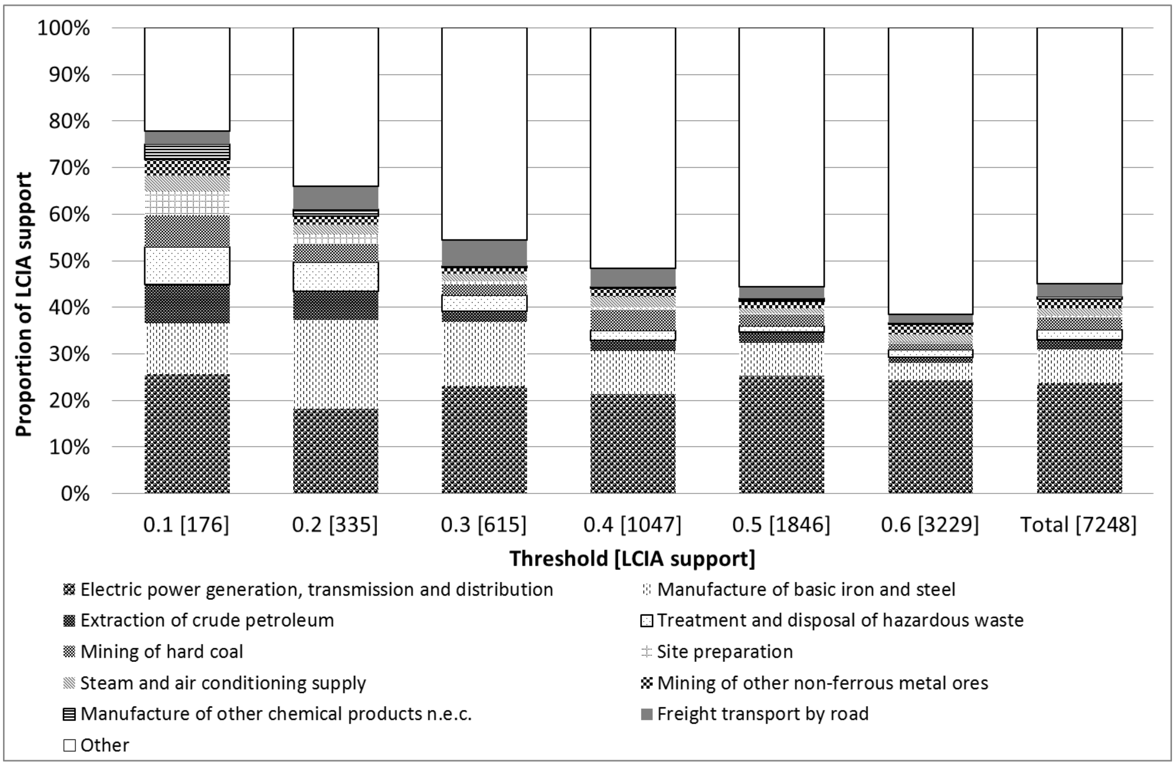


Figure S3-14: Accumulated LCIA support per threshold and ISIC rev. class for the connector perspective. The total LCIA support per threshold is given in square brackets. For clarity, only 9 ISIC categories are shown. The remaining 35 categories are consolidated into the "Other„ category.

Processes related to electric power generation, manufacture of basic iron and steel, treatment and disposal of hazardous waste, extraction of crude petroleum, mining of hard coal and freight transport are of highest importance. They accumulate roughly 70% of the total LCIA support in the first threshold.

3.7.2.7 The most important locations

Figure S10 shows the proportion of LCIA support per threshold differentiated into geographical locations.

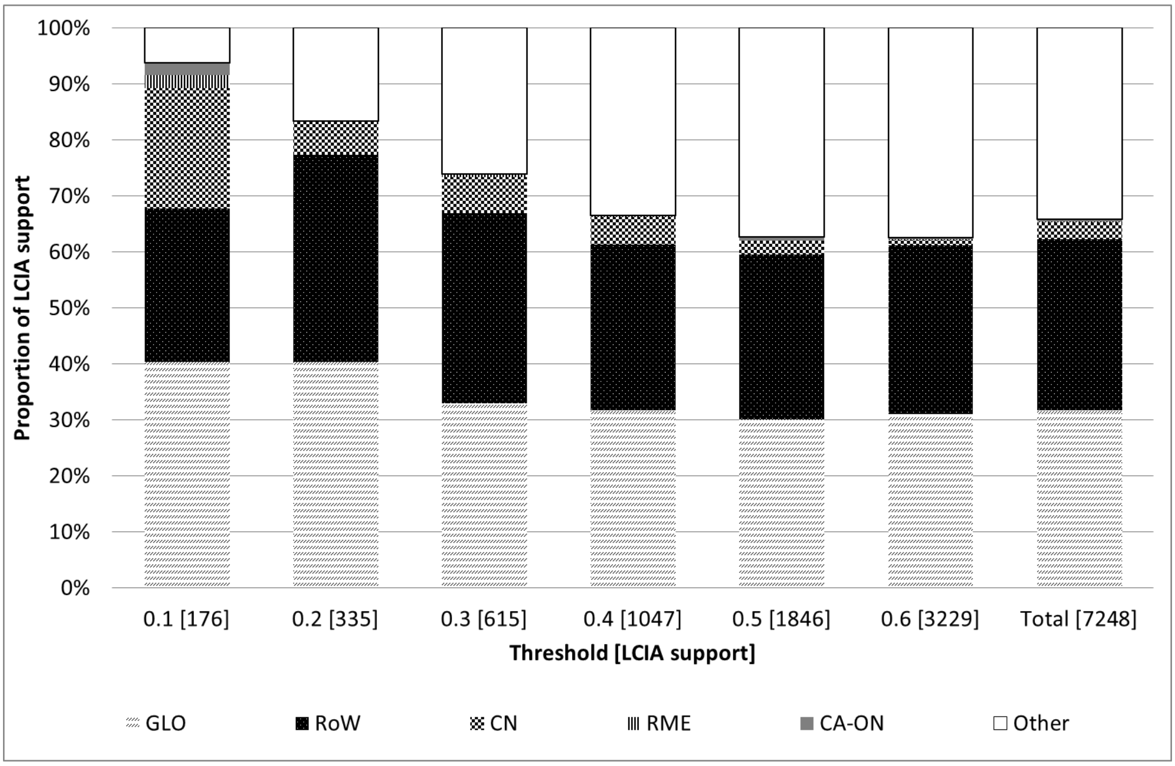


Figure S3-15: Accumulated LCIA support per threshold and geographical location. The total LCIA support per threshold is given in square brackets. Abbreviations: RoW= Rest-of-the-World; CN = China; GLO = Global; RME=Europe; CA-ON=Canada. The remaining 44 geographical locations are consolidated under "Other".

The connector perspective is dominated by processes referring to RoW, GLO and CN (see SI). Processes referring to these geographies cause almost 90% of the LCIA support in the first threshold, i.e., 27%, 22%, 40% for RoW, CN and GLO, respectively. Overall, datasets belonging to these geographies generate about 62% of the total LCIA support up to the sixth threshold.

3.7.3 Discussion

3.7.3.1 Inequality in the mean contribution vector and its source

Table S3-7 shows, for each of the assessed LCIA indicators, the (i) amount of characterization factors (CFs), (ii) the inventory support (IS), i.e. the amount of unit processes which include an elementary flow addressing one of the CFs, (iii) hx%, i.e. the amount of processes required to exceed a threshold of x%, (iv) the Gini coefficient (GC) for the causer and connector perspective. We use the GC in order to measure the inequality in the mean contribution vector associated with particular LCIA indicator (see Table S3-4, formula 19). The GC can take values between 0, when all mean contributions are equal and 1 when every unit process except one has a mean contribution of zero (Damgaard and Weiner, 2000).

Table S3-7: Selected summary statistics for the assessed set of LCIA indicators. Abbreviations: CF, characterization factors; IS, inventory support; hx%, amount of processes required to exceed a threshold of x%, GC, gini coefficient.

No.	LCIA indicator	# CFs	IS	h40%	h60%	GC causer	GC causer without zeros (non-contributing processes are excluded)	GC connector
1	CC	63	3'501	31	91	0.964	0.884	0.766
2	ARD	109	368	5	11	0.997	0.916	0.831
3	POF	154	3'369	20	70	0.963	0.877	0.772
4	MEP	29	2'678	19	73	0.962	0.839	0.765
5	IR	59	386	1	2	1.000	0.986	0.899
6	ODP	22	232	3	7	0.998	0.915	0.842
7	FEP	11	990	2	2	0.998	0.972	0.852
8	ALOP	17	342	5	9	0.997	0.896	0.854
9	TAF	11	2'559	17	51	0.973	0.880	0.779
10	HtoxC	216	2'849	3	5	0.996	0.986	0.842
11	HtoxNC	481	2'987	5	17	0.985	0.949	0.818
12	Etox	700	3'155	3	5	0.992	0.971	0.818
13	I+R	39	248	7	13	0.997	0.858	0.837
14	I+EQ	476	4'299	9	29	0.975	0.936	0.796
15	I+HH	550	4'036	21	61	0.966	0.908	0.776
16	I+CC	62	3'501	29	83	0.967	0.893	0.767
17	ES13_total	802	6'035	59	185	0.928	0.865	0.737
18	ES13Water	5	2'124	26	74	0.976	0.870	0.774
19	ReCiPe	899	5'132	37	109	0.954	0.900	0.765

The inventory support differs substantially across LCIA indicators (see also Figure S3-8). Among the assessed mid-point methods, LCIA indicators related to climate change have the largest inventory support – 3’500 processes cause emission with a global warming potential. In contrast, other midpoint methods such like ozone depletion or agricultural land occupation have an inventory support of 232 and 342, respectively. The Swiss Ecological Scarcity method (2013) is the endpoint method with the largest inventory support (6’035).

The GC confirms the high inequality in the size classes of the mean input contributions and indicates a high utility for prioritization across all LCIA indicators. $h_{60\%}$ and $h_{80\%}$ reveal how many processes we have to review to exceed a threshold of 40% and 60% for the particular LCIA indicator, respectively.

Table S3-8 shows a correlation analysis for selected metrics of Table S3-7.

Table S3-8: Correlation analysis. Abbreviations: Corr, correlation; CFs, characterization factors; GC, gini coefficient; IS, inventory support. Asterisks indicate the level of significance: *p<0.05, **p<0.01.

No.	Correlation metric	Pearson linear correlation coefficient		Kendall's tau		Spearman's rho	
	Parameter	**Corr**	**p**	**Corr**	**p**	**Corr**	**p**
1	**#CFs vs. IS**	0.747	2.10E-04**	0.524	2.10E-03**	0.737	3.22E-04**
2	**#CFs vs. GC causer**	-0.489	3.38E-02*	-0.282	9.99E-02	-0.400	8.14E-02
3	**IS vs. GC causer**	-0.852	3.57E-06**	-0.587	5.28E-04**	-0.793	5.23E-05**
4	**IS vs. GC causer without non-contributing processes**	-0.185	4.49E-01	-0.059	7.53E-01	-0.092	7.08E-01
5	**IS vs. GC_connector**	-0.779	8.51E-05**	-0.540	1.40E-03**	-0.7117	6.32E-04**
6	**GC_causer & GC_connector**	0.909	7.05E-08**	0.813	2.38E-08**	0.939	4.57E-06**

The amount of characterization factors per LCIA indicator (No. 1) seems to be a good predictor for the inventory support but shows only a minor (negative) correlation with the final inequality in the mean contribution vector (2). However, the rather large negative correlation between the inventory support and the final inequality (3) suggests that the inventory support (or lack therefore) is one important driver of the inequality in the mean contribution vector. This is confirmed by fact that the exclusion of non-contributing processes cancels the correlation between the inventory support and the inequality (4). Moreover, the inequality in the connector perspective seems to be largely associated with the inequality in the causer perspective (5).

3.7.3.2 References

Ahmadi Achachlouei, M., Hilty, L.M., 2015. Modeling the Effects of ICT on Environmental Sustainability: Revisiting a System Dynamics Model Developed for the European Commission, in: Hilty, M.L., Aebischer, B. (Eds.), ICT Innovations for Sustainability. Springer International Publishing, Cham, pp. 449–474.

Blonk, C., 2014. Agri-Footprint - Part 2 - Description of data - Version 1.0.

Bornhöft, N.A., Sun, T.Y., Hilty, L.M., Nowack, B., 2016. A dynamic probabilistic material flow modeling method. Environ. Model. Softw. 76, 69–80. https://doi.org/10.1016/j.envsoft.2015.11.012

Bourgault, G., Lesage, P., Samson, R., 2012. Systematic disaggregation: a hybrid LCI computation algorithm enhancing interpretation phase in LCA. Int. J. Life Cycle Assess. 17, 774–786. https://doi.org/10.1007/s11367-012-0418-7

Bulmer, M.G., 1979. Principles of Statistics. Dover Publications, Inc., New York.

Carlson, K.M., Gerber, J.S., Mueller, N.D., Herrero, M., MacDonald, G.K., Brauman, K.A., Havlik, P., O[rsquor]Connell, C.S., Johnson, J.A., Saatchi, S., West, P.C., 2016. Greenhouse gas emissions intensity of global croplands. Nat. Clim Change advance online publication.

Carlson, R., Tillman, A.-M., Steen, B., Löfgren, G., 1998. Lci data modelling and a database design. Int. J. Life Cycle Assess. 3, 106–113. https://doi.org/10.1007/BF02978500

Ciroth, A., 2013. OpenLCA - Process contribution - Documentation [WWW Document]. URL http://openlca.org/documentation/index.php/Process_contributions (accessed 8.16.13).

Ciroth, A., Fleischer, G., Steinbach, J., 2004. Uncertainty calculation in life cycle assessments. Int. J. Life Cycle Assess. 9, 216–226. https://doi.org/10.1007/BF02978597

Ciroth, A., Muller, S., Weidema, B., Lesage, P., 2013. Empirically based uncertainty factors for the pedigree matrix in ecoinvent. Int. J. Life Cycle Assess. 1–11. https://doi.org/10.1007/s11367-013-0670-5

Damgaard, C., Weiner, J., 2000. DESCRIBING INEQUALITY IN PLANT SIZE OR FECUNDITY. Ecology 81, 1139–1142. https://doi.org/10.1890/0012-9658(2000)081[1139:DIIPSO]2.0.CO;2

Davis, C., Nikolic, I., Dijkema, G.P.J., 2010. Industrial Ecology 2.0. J. Ind. Ecol. 14, 707–726. https://doi.org/10.1111/j.1530-9290.2010.00281.x

DMIC, 2017. ::: Welcome to Delhi Mumbai Industrial Corridor ::: [WWW Document]. URL http://delhimumbaiindustrialcorridor.com/ (accessed 1.18.17).

Dresen, B., Jandewerth, M., 2012. Integration of spatial analyses into LCA—calculating GHG emissions with geoinformation systems. Int. J. Life Cycle Assess. 17, 1094–1103. https://doi.org/10.1007/s11367-011-0378-3

Duclos, J.-Y., Araar, A., 2006. Poverty and Equity: Measurement, Policy, and Estimation with DAD. Springer.

Düpmeier, C., Greceanu, C., Kusche, O., Schmitt, C., 2012. Managing LCI Data from Different Workgroups within the same Instance of an LCA Database., in: EnviroInfo Dessau 2012, Dessau, Germany, August 29-31, 2012. pp. 507–515.

Earth Observation Center, 2017. . Copernic. Eur. Erdbeobachtungsprogramm. URL http://www.dlr.de/eoc/desktopdefault.aspx/tabid-5367/9013_read-16792/

Ebner, F., 2013. Results, Interpretation and Reporting: GaBi Software [WWW Document]. Gabi-Softw. URL http://www.gabi-software.com/software/gabi-software/gabi-5/functionalities/results-interpretation-and-reporting/ (accessed 8.16.13).

EC-JRC, 2011. International Reference Life Cycle Data System (ILCD) Handbook-Recommendations for Life Cycle Impact Assessment in the European context. First edition November 2011. EUR 24571 EN. Luxemburg. Publications Office of the European Union;

Ecoinvent, 2015. Allocation cut-off by classification [WWW Document]. URL http://www.ecoinvent.org/database/ecoinvent-version-3/system-models/allocation-cut-off-by-classification/ (accessed 4.16.15).

European Commission, 2009. Directive 2009/125/EC of the European Parliament and of the Council of 21 October 2009 establishing a framework for the setting of ecodesign requirements for energy-related products.

FAO, 2017. Agro-Maps. Global Spatial Database of Agricultural Land-use Statistics [WWW Document]. URL http://kids.fao.org/agromaps/

FAOSTAT, 2016. . Prod. Crops. URL http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor

Finnveden, G., Hauschild, M.Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., Suh, S., 2009. Recent developments in Life Cycle Assessment. J. Environ. Manage. 91, 1–21. https://doi.org/10.1016/j.jenvman.2009.06.018

Finnveden, G., Moberg, Å., 2005. Environmental systems analysis tools – an overview. J. Clean. Prod. 13, 1165–1173. https://doi.org/10.1016/j.jclepro.2004.06.004

Frischknecht, R., Althaus, H.-J., Bauer, C., Doka, G., Heck, T., Jungbluth, N., Kellenberger, D., Nemecek, T., 2007. The environmental relevance of capital goods in life cycle assessments of products and services. Int. J. Life Cycle Assess. 12, 7–17.

Gerber, J.S., Carlson, K.M., Makowski, D., Mueller, N.D., Garcia de Cortazar-Atauri, I., Havlík, P., Herrero, M., Launay, M., O'Connell, C.S., Smith, P., West, P.C., 2016. Spatially explicit estimates of N2O emissions from croplands suggest climate mitigation opportunities from improved fertilizer management. Glob. Change Biol. 22, 3383–3394. https://doi.org/10.1111/gcb.13341

Geyer, R., Lindner, J.P., Stoms, D.M., Davis, F.W., Wittstock, B., 2010a. Coupling GIS and LCA for biodiversity assessments of land use. Int. J. Life Cycle Assess. 15, 692–703. https://doi.org/10.1007/s11367-010-0199-9

Geyer, R., Stoms, D., Lindner, J., Davis, F., Wittstock, B., 2010b. Coupling GIS and LCA for biodiversity assessments of land use. Int. J. Life Cycle Assess. 15, 454–467. https://doi.org/10.1007/s11367-010-0170-9

Goedkoop, M., Oele, M., 2004. SimaPro 6 - Introduction to LCA with SimaPro.

Gregor, S., Hevner, A.R., 2013. Positioning and Presenting Design Science Research for Maximum Impact. MIS Q. 37, 337–355.

Groen, E.A., Heijungs, R., Bokkers, E.A.M., de Boer, I.J.M., 2014. Methods for uncertainty propagation in life cycle assessment. Environ. Model. Softw. 62, 316–325. https://doi.org/10.1016/j.envsoft.2014.10.006

Guinée, J.B., Gorrée, M., Heijungs, R., Huppes, G., Kleijn, R., de Koning, A., van Oers, L., Sleeswijk, A.W., Suh, S., de Haes, H.A.U., de Bruijn, H., van Duin, R., Huijbregts, M.A.J., 2002. Handbook on Life Cycle Assessment. Operational Guide to the ISO Standards.

Guinée, J.B., Heijungs, R., Huppes, G., Zamagni, A., Masoni, P., Buonamici, R., Ekvall, T., Rydberg, T., 2011. Life Cycle Assessment: Past, Present, and Future. Environ. Sci. Technol. 45, 90–96. https://doi.org/10.1021/es101316v

Hauschild, M., 2006. Spatial Differentiation in Life Cycle Impact Assessment: A decade of method development to increase the environmental realism of LCIA. Int. J. Life Cycle Assess. 11, 11–13. https://doi.org/10.1065/lca2006.04.005

Heijungs, R., 2015. Topological Network Theory and Its Application to LCA and IOA and Related Industrial Ecology Tools: Principles and Promise. J. Environ. Acccounting Manag. Volume 3, 151–167. https://doi.org/10.5890/JEAM.2015.06.005

Heijungs, R., 2012. Spatial differentiation, GIS-based regionalization, hyperregionalization, and the boundaries of LCA., in: Environment and Energy (Editorial Series of Italian Commodity Science Academy and Engineering Association of Messina). FrancoAngeli, Milano, Italy, pp. 165–176.

Heijungs, R., 1996. Identification of key issues for further investigation in improving the reliability of life-cycle assessments. J. Clean. Prod. 4, 159–166. https://doi.org/10.1016/S0959-6526(96)00042-X

Heijungs, R., Huijbregts, M., 2004. A review of approaches to treat uncertainty in LCA. Complex Integr Resour Manag Proc 2nd Bienn Meet Int Env. Model Softw Soc.

Heijungs, R., Kleijn, R., 2001. Numerical approaches towards life cycle interpretation five examples. Int. J. Life Cycle Assess. 6, 141–148. https://doi.org/10.1007/BF02978732

Heijungs, R., Suh, S., 2002. The computational structure of life cycle assessment. Kluwer Academic Publishers, Dordrecht.

Heijungs, R., Suh, S., Kleijn, R., 2005. Numerical Approaches to Life Cycle Interpretation - The case of the Ecoinvent'96 database (10 pp). Int. J. Life Cycle Assess. 10, 103–112. https://doi.org/10.1065/lca2004.06.161

Hellweg, S., Milà i Canals, L., 2014. Emerging approaches, challenges and opportunities in life cycle assessment. Science 344, 1109–1113.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — Global Soil Information Based on Automated Mapping. PLoS ONE 9, e105992. https://doi.org/10.1371/journal.pone.0105992

Hevner, A., Chatterjee, S., 2010. Design Research in Information Systems: Theory and Practice. Springer.

Huijbregts, M.J., 1998. Application of uncertainty and variability in LCA. Int. J. Life Cycle Assess. 3, 273–280. https://doi.org/10.1007/BF02979835

IIASSA/FAO, 2012. Global Agro-ecological Zones (GAEZ v3.0). IIASA, Laxenburg / FAO, Rome.

ISO, 2006. ISO 14044. Environmental management - Life cylce assessment - Requirements and guidelines.

Kostadinov, F., Holm, S., Steubing, B., Thees, O., Lemm, R., 2014. Simulation of a Swiss wood fuel and roundwood market: An explorative study in agent-based modeling. For. Policy Econ. 38, 105–118. https://doi.org/10.1016/j.forpol.2013.08.001

Lenzen, M., 2000. Errors in Conventional and Input-Output—based Life—Cycle Inventories. J. Ind. Ecol. 4, 127–148. https://doi.org/10.1162/10881980052541981

Lenzen, M., Dey, C., 2000. Truncation error in embodied energy analyses of basic iron and steel products. Energy 25, 577–585. https://doi.org/10.1016/S0360-5442(99)00088-2

Lesage, P., Samson, R., 2013. The Quebec Life Cycle Inventory Database Project. Int. J. Life Cycle Assess. 1–8. https://doi.org/10.1007/s11367-013-0593-1

Majeau-Bettez, G., Strømman, A.H., Hertwich, E.G., 2011. Evaluation of Process- and Input–Output-based Life Cycle Inventory Data with Regard to Truncation and Aggregation Issues. Environ. Sci. Technol. 45, 10170–10177. https://doi.org/10.1021/es201308x

Monfreda, C., Ramankutty, N., Foley, J.A., 2008. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Glob. Biogeochem. Cycles 22. https://doi.org/10.1029/2007GB002947.

Mueller, N.D., Gerber, J.S., Johnston, M., Ray, D.K., Ramankutty, N., Foley, J.A., 2012. Closing yield gaps through nutrient and water management. Nature 490, 254–257. https://doi.org/10.1038/nature11420

Mutel, C., 2015. A new open source framework for advanced life cycle assessment calculations [WWW Document]. URL https://brightwaylca.org/

Mutel, C., 2012. Framework and Tools for Regionlization in Life Cycle Assessment. ETH, Zurich.

Mutel, C.L., Hellweg, S., 2009. Regionalized Life Cycle Assessment: Computational Methodology and Application to Inventory Databases. Environ. Sci. Technol. 43, 5797–5803. https://doi.org/10.1021/es803002j

Mutel, C.L., Pfister, S., Hellweg, S., 2012. GIS-based regionalized life cycle assessment: how big is small enough? Methodology and case study of electricity generation. Environ. Sci. Technol. 46, 1096–1103. https://doi.org/10.1021/es203117z

Natural Earth, 2016. Admin 0 - Country data [WWW Document]. Nat. Earth. URL ww.naturalearthdata.com/http//www.naturalearthdata.com/download/110m/cultural/ne_110m_admin_0_countries_lakes.zip

Nemecek, T., Bengoa, X., Lansche, J., Mouron, P., Riedener, E., Rossi, V., Humbert, S., 2015. Methodological Guidelines for the Life Cycle Inventory of Agricultural Products. World Food LCA Database (WFLDB), Version 3.0.

Nemecek, T., Schnetzer, J., Reinhard, J., 2014. Updated and harmonised greenhouse gas emissions for crop inventories. Int. J. Life Cycle Assess. 1–18. https://doi.org/10.1007/s11367-014-0712-7

Panagos, P., Borelli, P., Meusburger, K., 2015. A New European Slope Length and Steepness Factor (LS-Factor) for Modeling Soil Erosion by Water. Geosciences 5, 117–126.

Panagos, P., Meusburger, K., Ballabio, C., Borrelli, P., Alewell, C., 2014. Soil erodibility in Europe: A high-resolution dataset based on LUCAS. Sci. Total Environ. 479–480, 189–200. https://doi.org/10.1016/j.scitotenv.2014.02.010

Peffers, K., Tuunanen, T., Rothenberger, M., Chatterjee, S., 2007. A Design Science Research Methodology for Information Systems Research. J Manage Inf Syst 24, 45–77.

Pfister, S., Bayer, P., Koehler, A., Hellweg, S., 2011. Environmental Impacts of Water Use in Global Crop Production: Hotspots and Trade-Offs with Land Use. Environ. Sci. Technol. 45, 5761–5768. https://doi.org/10.1021/es1041755

Potter, P., Ramankutty, N., Bennett, E.M., Donner, S.D., 2011a. Global Fertilizer and Manure, Version 1: Nitrogen in Manure Production.

Potter, P., Ramankutty, N., Bennett, E.M., Donner, S.D., 2011b. Global Fertilizer and Manure, Version 1: Phosphorus in Manure Production.

Potter, P., Ramankutty, N., Bennett, E.M., Donner, S.D., 2010. Characterizing the Spatial Patterns of Global Fertilizer Application and Manure Production. Earth Interact. 14, 1–22.

Read L., L., 1958. I, pencil.

Rebitzer, G., Ekvall, T., Frischknecht, R., Hunkeler, D., Norris, G., Rydberg, T., Schmidt, W.-P., Suh, S., Weidema, B.P., Pennington, D.W., 2004. Life cycle assessment: Part 1: Framework, goal and scope definition, inventory analysis, and applications. Environ. Int. 30, 701–720. https://doi.org/10.1016/j.envint.2003.11.005

Reinhard, J., Emmenegger, M.F., Widok, A.H., Wohlgemuth, V., 2011. RSB tool: a LCA tool for the assessment of biofuels sustainability, in: Proceedings of the Winter Simulation Conference. Winter Simulation Conference, Phoenix, Arizona, pp. 1048–1059.

Reinhard, J., Mutel, C., Wernet, G., Zah, R., Hilty, L., 2016. Contribution-based prioritization of LCI database improvements: Method design, demonstration, and evaluation. Environ. Model. Softw. 86, 204–218. https://doi.org/10.1016/j.envsoft.2016.09.018

Rørbech, J.T., Vadenbo, C., Hellweg, S., Astrup, T.F., 2014. Impact Assessment of Abiotic Resources in LCA: Quantitative Comparison of Selected Characterization Models. Environ. Sci. Technol. 48, 11072–11081. https://doi.org/10.1021/es5023976

Scherer, L., Pfister, S., 2015. Modelling spatially explicit impacts from phosphorus emissions in agriculture. Int. J. Life Cycle Assess. 20, 785–795. https://doi.org/10.1007/s11367-015-0880-0

Seto, K.C., Reenberg, A., Boone, C.G., Fragkias, M., Haase, D., Langanke, T., Marcotullio, P., Munroe, D.K., Olah, B., Simon, D., 2012. Urban land teleconnections and sustainability. Proc. Natl. Acad. Sci.

Slottje, D., Nieswiadomy, M., Redfearn, M., 2001. Economic inequality and the environment. Environ. Model. Socioecon. 16, 183–194. https://doi.org/10.1016/S1364-8152(00)00081-5

Sonnemann, G., Vigon, B. (eds)., 2011. Global Guidance Principles for Life Cycle Assessment Databases: A Basis for Greener Processes and Products (Shonan Guidance Principles).

Steubing, B., Wernet, G., Reinhard, J., Bauer, C., Moreno-Ruiz, E., 2016. The ecoinvent database version 3 (part II): analyzing LCA results and comparison to version 2. Int. J. Life Cycle Assess. 21, 1269–1281. https://doi.org/10.1007/s11367-016-1109-6

Suh, S., Huppes, G., 2005. Methods for Life Cycle Inventory of a product. J. Clean. Prod. 13, 687–697. https://doi.org/10.1016/j.jclepro.2003.04.001

Suh, S., Huppes, G., 2002. Missing inventory estimation tool using extended input-output analysis. Int. J. Life Cycle Assess. 7, 134–140. https://doi.org/10.1007/BF02994047

Tillman, A.-M., 2000. Significance of decision-making for LCA methodology. Environ. Impact Assess. Rev. 20, 113–123. https://doi.org/10.1016/S0195-9255(99)00035-9

USGS, 2016. Mineral Resources Data System (MRDS) [WWW Document]. URL https://mrdata.usgs.gov/mrds/

Venable, J., Pries-Heje, J., Baskerville, R., 2012. A Comprehensive Framework for Evaluation in Design Science Research, in: Peffers, K., Rothenberger, M., Kuechler, B. (Eds.), Design Science Research in Information Systems. Advances in Theory and Practice, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 423–438.

Weidema, B., Bauer, C., Hischier, R., Mutel, C., Nemecek, T., Reinhard, J., Vadenbo, C., Wernet, G., 2013. Overview and methodology. Data quality guideline for the ecoinvent database version 3.

Werner, F., 2002. Interdependencies between LC-Modelling and the use of LCA in product design-related decision situations. With special emphasis on the influence of cognitive models and values on the modelling of reuse & recycling and other end-of-life options (Diss. ETH Nr. 14‘750). Zurich.

Wernet, G., 2012. ecoinvent v3 - Internationalization of the data supply and international collaborations. Presentation at DF48, Dübendorf.

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno, E., Weidema, B., submitted. The ecoinvent database version 3 (part I): overview and methodology.

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., Weidema, B., 2016. The ecoinvent database version 3 (part I): overview and methodology. Int. J. Life Cycle Assess. 21, 1218–1230. https://doi.org/10.1007/s11367-016-1087-8

YANG, X.. P., W.M...THORNTON, P.E...JAIN, A.K.., 2014. Global Gridded Soil Phosphorus Distribution Maps at 0.5-degree Resolution.

Zah, R., Faist, M., Reinhard, J., Birchmeier, D., 2009. Standardized and simplified life-cycle assessment (LCA) as a driver for more sustainable biofuels. J. Clean. Prod. 17, Supplement 1, S102–S105. https://doi.org/10.1016/j.jclepro.2009.04.004

Zah, R., Gmuender, S., Gauch, M., 2012. Evaluación del ciclo de vida de la cadena de producción de biocombustibles en Colombia (No. Estrategias de energía sostenible y biocombustibles para Colombia ATN/JC-10826-CO y ATN/JF-10827-CO). Banco Interamericano de Desarrollo (BID).

4 Article III: Regionalized LCI Modeling. A Framework for the integration of spatial data in Life Cycle Assessment

Juergen Reinhard, Rainer Zah and Lorenz M. Hilty

Status: *Published in "Wohlgemuth, V., Fuchs-Kittowski, F., Wittmann, J. (Eds.), Advances and New Trends in Environmental Informatics: Stability, Continuity, Innovation. Springer International Publishing, Cham, pp. 223–235."*

Abstract Life Cycle Assessment (LCA), the most prominent technique for the assessment of environmental impacts of products, typically operates on the basis of average meteorological and ecological conditions of whole countries or large regions. This limits the representativeness and accuracy of LCA, particularly in the field of agriculture. The production processes associated with agricultural commodities are characterized by high spatial sensitivity as both inputs (e.g. mineral and organic fertilizers) and the accompanying release of emissions into soil, air and water (e.g. nitrate, dinitrogen monoxide, or phosphate emissions) are largely determined by micro-spatial environmental parameters (precipitation, soil properties, slope, etc.) and therefore highly context dependent. This spatial variability is vastly ignored under the "unit world" assumption inherent to LCA. In this paper, we present a new calculation framework for regionalized life cycle inventory modeling that aims to overcome this inherent limitation. The framework allows an automated, site-specific generation and assessment of regionalized unit process datasets. We demonstrate the framework in a case study on rapeseed cultivation in Germany. The results from the research are (i) a framework for generating regionalized data structures, and (ii) a first examination of the significance of further use cases.

Keywords Regionalization • Site-specific LCI modeling • LCA • Raster data

4.1 Introduction

In order to address the challenges associated with climate change and other environmental threats, environmental considerations need to be integrated in many types of decisions (Finnveden et al., 2009). This highlights the importance of methods and tools for measuring and comparing the environmental impacts of human activities for the provision of goods and services. One of the most prominent methods in this regard is Life Cycle Assessment (LCA). LCA is a technique for the comprehensive quantitative assessment of the environmental impacts of products in a life-cycle perspective (Finnveden et al., 2009). It focuses on the compilation and environmental evaluation of all inputs and outputs associated with a product throughout its life cycle with the goal to pinpoint ecological weaknesses, compare possible alternatives, evaluate the main impact factors, design new products, measure the environmental relevance of a material or product and establish recommendations for actions (Guinée et al., 2002).

The environmental impact of a product is caused by the exchange of energy and matter between its technical (product) system and its surrounding environment (Carlson et al., 1998). The technical system is represented as a linear model composed of a network of nodes called unit process datasets (UPDs) (Carlson et al., 1998). Each UPD includes data on (i) the *intermediate exchanges*, i.e., the input of energy and material flows and the output of products and waste flows, and (ii) the *exchanges with environment*, i.e., the inputs of natural resources and outputs of emissions (Carlson et al., 1998). The UPDs are linearly linked via their intermediate exchanges. The calculation of the product system for the reference flow (product) of interest reveals the throughput of all exchanges with environment, the so-called Life Cycle Inventory (LCI). Life Cycle Impact Assessment (LCIA) – the succeeding step of the LCA procedure – then takes the inventory data on these exchanges as an input to determine the impacts on the surrounding environment.

It is common practice to consider average meteorological and ecological conditions of a whole country and geographical region for the compilation of the intermediate exchanges and exchanges with environment listed in the UPDs. That is, LCA in practice is mostly structured around the use of either *site-generic*[50] or *site-dependent*[51] unit process datasets (UPD), whereas *site-specific*[52] data is used very rarely. The high cost of primary data collection only allows for rudimentary site-specific UPD generation. That is, LCA usually must focus on average values at the expense of specificity (Mutel and Hellweg, 2009).

While this approach is suitable for determining important key drivers of the environmental impacts, it limits the range of questions which can be properly addressed with the LCA technique. For example, UPDs representing agricultural cultivation comprise exchange flows characterized by

[50] Site-generic values represent an average over large geographic regions, such as continents or the globe (Mutel et al., 2012).

[51] Site-dependent values follow country or state boundaries (Mutel et al., 2012).

[52] Site-specific values are usually used only for individual locations, such as a particular plot, factory or landfill (Mutel et al., 2012).

a high sensitivity to the natural variability of the surrounding environment. The type and amount of resources used (e.g. water, land), the intermediate flows required (e.g. the application of mineral and organic fertilizer or the use of machinery) and the accompanying release of emissions into soil, air and water (e.g. nitrate, di-nitrogen monoxide, phosphate emissions) are determined by micro-spatial environmental parameters (precipitation, soil properties, slope, etc.) and therefore highly context dependent. That is, for such processes even small changes in local, bio-geographical conditions can alter the type and magnitude of the included exchange flows and hence also their environmental impacts (Geyer et al., 2010a).

Regionalized LCI modeling is motivated by the "recognition that industrial production characteristic vary throughout space" (Mutel et al., 2012). We believe that the development of a computerized technique for spatially explicit (regionalized) LCI modeling in LCA would decrease the costs of collecting and computing more and better information about agricultural production characteristics. The potential of regionalized *LCI* modeling has already been acknowledged (Mutel et al., 2012; Mutel and Hellweg, 2009; Seto et al., 2012) although most of the recent literature on regional aspects in LCA focuses on regionalized LCIA modeling –recognizing that the location of a source and the conditions of its surroundings influence the environmental impact (Hauschild, 2006). To the best of our knowledge, regionalized LCI modeling is to date only applied within six case studies (Dresen and Jandewerth, 2012; Geyer et al., 2010a, 2010b; Reinhard et al., 2011; Scherer and Pfister, 2015; Zah et al., 2012). All of these studies succeed in the parameterization of the spatial properties of selected exchanges. However, none of the studies analyzed provides a *general framework* for regionalized LCI modeling in LCA.

In this paper, we present a prototype calculation framework[53] that allows the automated, site-specific (regionalized) generation and assessment of cradle-to-gate agricultural UPDs. We transform publicly available spatial (raster) data (Table 4-1) into comprehensive UPDs using default data from version 3.2 of the ecoinvent database and the emission models from the World Food Life Cycle Database (WFLDB) Guidelines (Nemecek et al., 2015). We present a case study of rapeseed production in Germany to illustrate the framework. Using a resolution of 30 arc seconds (~1 x 1 km), we generate and assess roughly 580'000 regionalized UPDs for rapeseed cultivation in Germany. We conclude with a discussion of limitations and further use cases of the presented framework.

[53] The framework is currently still under development and therefore not publicly available.

4.2 Method

4.2.1 Framework for regionalized LCI modeling

We compiled a repository of publicly available raster data (see Table 4-1) indicating harvested area, yield, fertilizer application rates, irrigation requirement of all major crops as well as data on precipitation, soil properties and terrain. Almost all of the data has a global scale (see Table 4-1). This ensures that the framework is applicable in all regions of the world.

Table 4-1: The compiled repository of spatial raster datasets.

(No.) Institution	Content	Resolution	Spatial and temporal extent	Source
(1) Global Agro-Ecological Zones (GAEZ)	Precipitation	5*5 minute	World, mean, 1961 - 1990	(IIASSA/FAO, 2012)
(2) Global Agro-Ecological Zones (GAEZ)	Length of crop growing cycle, rapeseed	5*5 minute	World, mean, 1961 - 1990	(IIASSA/FAO, 2012)
(3) Publication	Irrigation requirement	Country	World, 2000	(Pfister et al., 2011)
(4) ISRIC, Soil grids	Soil property data (pH, coarse fragments, organic carbon content, bulk density, clay content)	30*30 arc second	World, 2014	(Hengl et al., 2014)
(5) Earth Stat	Fertilizer application (N, P & K) for major crops	5*5 minute	World, 2000	(Mueller et al., 2012)
(6) Earth Stat	Harvested area and yield for 175 Crops	5*5 minute	World, 2000	(Monfreda et al., 2008)
(7) Earth data	P content of soils	0.5*0.5 degree	World, 2014	(YANG, 2014)
(8) Earth data	Amount of nitrogen manure produced and present on the landscape	0.5*0.5 degree	World, 2001	(Potter et al., 2010, 2011a)
(9) Earth data	Amount of phosphorous in manure produced and present on the landscape	0.5*0.5 degree	World, 2001	(Potter et al., 2010), (Potter et al., 2011b)
(10) European Soil Data Centre (ESDAC)	Soil erodibility factor	30*30 arc second	Europe, 2014	(Panagos et al., 2014)
(11) European Soil Data Centre (ESDAC)	Length slope factor	30*30 arc second	Europe, 2015	(Panagos et al., 2015)
(12) Natural Earth	Country shape files		World, 2015	(Natural Earth, 2016)

We process this repository into UPDs with the following procedure, implemented in python 3.4 (Figure 4-1):

1. We clip the raster data to the spatial extend (e.g. country borders, province borders, etc.) of interest. We use GDAL (Geospatial Data Abstraction Library) and shape files from natural earth for this task.

2. We load the crop and country specific raster data into data frames using Pandas, a Python Data Analysis Library. We merge[54] these data frames into one table. Each column of this parameter table represents, for a given latitude-longitude combination, the corresponding grid cell values of all clipped raster datasets (see Figure 4-1, parameter table). The parameter table serves as the basis for the further processing of the data.
3. We use the emission models from the WFLDB guidelines (Nemecek et al., 2015, p. 35 Table 5) to calculate new parameters, i.e., add rows to the parameter table. Each model takes several rows as input and produces one or more rows as output. For example, the calculation of nitrate emissions requires the parameter SOC, i.e., the soil organic carbon content (see step 3). We use equation 1 and the spatially explicit background data from ISRIC soil grids to compute SOC in kg C/ha for a soil depth of 0.3 m for each grid cell.

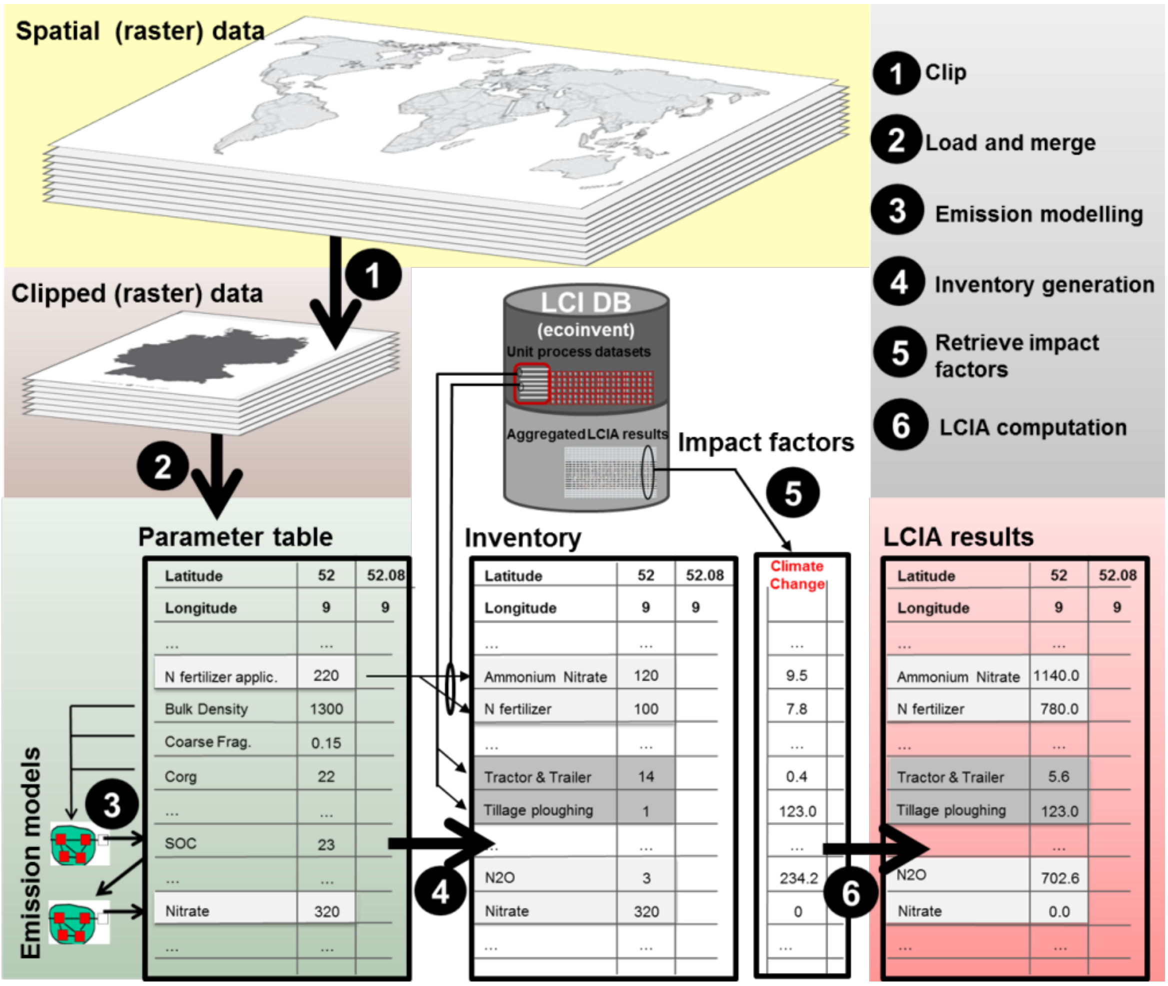
$$\mathrm{SOC} = \frac{C_{org}}{1000} * 3000 * BD * (1 - CF) \quad (1)$$

where: C_{org}= organic carbon content in soil in g/kg soil, BD = Bulk density in kg/m^3, CF = coarse fragments in %, 3000 represents the amount of m^3/ha and 1000 the transformation from g to kg/kg soil.

4. In this step (see 4 in Figure 4-1), we transform selected rows of the parameter table into a comprehensive inventory table that match the common agricultural UPD structure of the ecoinvent database. We access the UPD data on the basis of brightway2, an open source framework for advanced LCA (Mutel, 2015). The direct access to the ecoinvent database facilitates the matching and expansion of selected parameter table entries with a corresponding intermediate product or an elementary exchanges present in the ecoinvent database. For example, the N application rates from Earth Stat (Mueller et al., 2012) can be converted into the generic N-based mineral fertilizer products available in the ecoinvent database. Furthermore, exchange flows which cannot be computed from the parameter table due to the lack of regionalized data or models (e.g. transport tractor and trailer, mechanical field work) can be extrapolated on the basis of default data obtained from existing ecoinvent UPDs. Each column of the final inventory table then represents a regionalized UPD for a particular grid cell in ecoinvent nomenclature. That is, step 4 of the framework already yields regionalized cradle-to-gate agricultural UPDs.

[54] To date, the merging of the data is the most time-consuming step (see discussion section).

5. We generate a vector of impact factors for each exchange flow in the inventory table. We compute the aggregated LCIA impacts for each intermediate exchange (e.g. fertilizer, mechanical field work) and retrieve the characterization factor for each elementary exchange (e.g. N_2O emission). We can generate this vector for any LCIA indicator implemented in the ecoinvent database[55]. Again, we use the brightway2 framework for this task.
6. Finally, we assess the environmental impacts of the regionalized inventory. We multiply the LCIA impact factors for a particular row with the corresponding exchange flow value retrieved from the inventory table. The resulting LCIA table (see Figure 4-1, step 6) contains for each grid cell the LCIA impacts associated with each exchange flow value in the inventory table. The column sums of this table then represent the total environmental impacts of each grid cell for a particular LCIA indicator.



Latitude	52	52.08
Longitude	9	9
...	...	
N fertilizer applic.	220	
Bulk Density	1300	
Coarse Frag.	0.15	
Corg	22	
...	...	
SOC	23	
...	...	
Nitrate	320	
...	...	

Latitude	52	52.08
Longitude	9	9
...	...	
Ammonium Nitrate	120	
N fertilizer	100	
...	...	
Tractor & Trailer	14	
Tillage ploughing	1	
...	...	
N2O	3	
Nitrate	320	
...	...	

Climate Change
...
9.5
7.8
...
0.4
123.0
...
234.2
0
...

Latitude	52	52.08
Longitude	9	9
...	...	
Ammonium Nitrate	1140.0	
N fertilizer	780.0	
...	...	
Tractor & Trailer	5.6	
Tillage ploughing	123.0	
...	...	
N2O	702.6	
Nitrate	0.0	
...	...	

Figure 4-1: Our framework to transform spatial raster data into UPDs. A detailed explanation is given in the text.

[55] That is, the LCIA vector can be generated for 692 LCIA indicators.

We illustrate the application of this framework using a case study of rapeseed cultivation in Germany.

4.2.2 Case study of rapeseed cultivation

We generate and assess UPDs for a resolution of ~1 x 1 km, i.e., in the resolution of the soil property raster data[56]. We compute UPDs for each grid cell where rapeseed cultivation takes place. Gaps in raster data are excluded from the analysis. The exchange flows in each regionalized UPD refer to the cultivation of one hectare of rapeseed in a cradle-to-gate perspective[57].

We first generate the parameter table by clipping and merging all data in Table 4-1 to the geographical extent of Germany (step 1 & 2). In step 3, we compute Ammonia, Nitrous oxide, Dinitrogen monoxide, Nitrate and Phosphorus emissions according to the WFLDB guidelines (Nemecek et al., 2015, p. 35 Table 5).

We next generate the inventory table based on selected columns in the parameter table and default data from the German rapeseed cultivation dataset (step 4). Table 4-2 shows, for selected inventory flows, how we integrate such default data to generate an inventory table that match the common agricultural UPD structure of the ecoinvent database.

Table 4-2: Integration of default values from the ecoinvent database.

Parameter table	Inventory table	Computation	Comment / Source
N application rate (NAR)	Ammonium nitrate, as N	= NAR * 0.55	Fertilizer type and proportion is taken from the ecoinvent datasets “rapeseed production, DE”. Factors (2.3 and 1.2) transform from P to P_2O_5 and K to K_2O, respectively.
	Nitrogen fertilizer, as N	= NAR *0.45	
P application rate (PAR)	Phosphate fertilizer, as P_2O_5	= PAR * 2.3	
K application rate (KAR)	Potassium chloride, as K_2O	= KAR * 1.2	
Tractor & Trailer; Tillage; Ploughing & harrowing; Sawing; Fertilizing by broadcaster; Combine harvesting; Pesticide, unspecified.		= Value in generic ecoinvent dataset “rapeseed production, DE”	Taken by default from the ecoinvent dataset

[56] This means that due to the difference in resolution between the crop specific raster datasets (see no. 5 and no. 4 in Table 4-1) and the soil property raster datasets (see no. 4, 10 & 11 in Table 4-1) the 100 rapeseed UPDs generated for a ~10 x 10 km grid cell will always have the same yield and fertilizer input. That is, within a ~10 x 10 km grid cell, the only varying parameters are the soil properties and the therefrom computed emissions. Although this computation strategy comes at higher computational costs—for one parameter, we have to extract the values of roughly one million grid cells—it pays off by enabling us to assess the influence of changing soil properties on emissions in a ceteris paribus examination.

[57] That is, all upstream interventions are included. The further usage of the rapeseed (e.g. as feed or biofuel) is not considered.

After elimination of data gaps[58], the final inventory table contains roughly about 588’000 columns, i.e. UPDs. We apply three LCIA midpoint indicators to assess the environmental impacts for each UPD in the inventory table (step 5 & 6): climate change (CC, IPCC2013 GWP100a), marine eutrophication (MEP, ReCiPe Midpoint (H)) and freshwater eutrophication (FEP, ReCiPe Mipoint (H)).

4.3 Results

Figure 4-2 shows the spatially explicit environmental impacts per hectare rapeseed cultivated for CC, FEP and MEP. The spatial distributions of the environmental impacts differ depending on the LCIA indicators applied. While rapeseed cultivation in Lower Saxony and Schleswig-Holstein causes the highest CC impacts, both states show very low impacts for FEP.

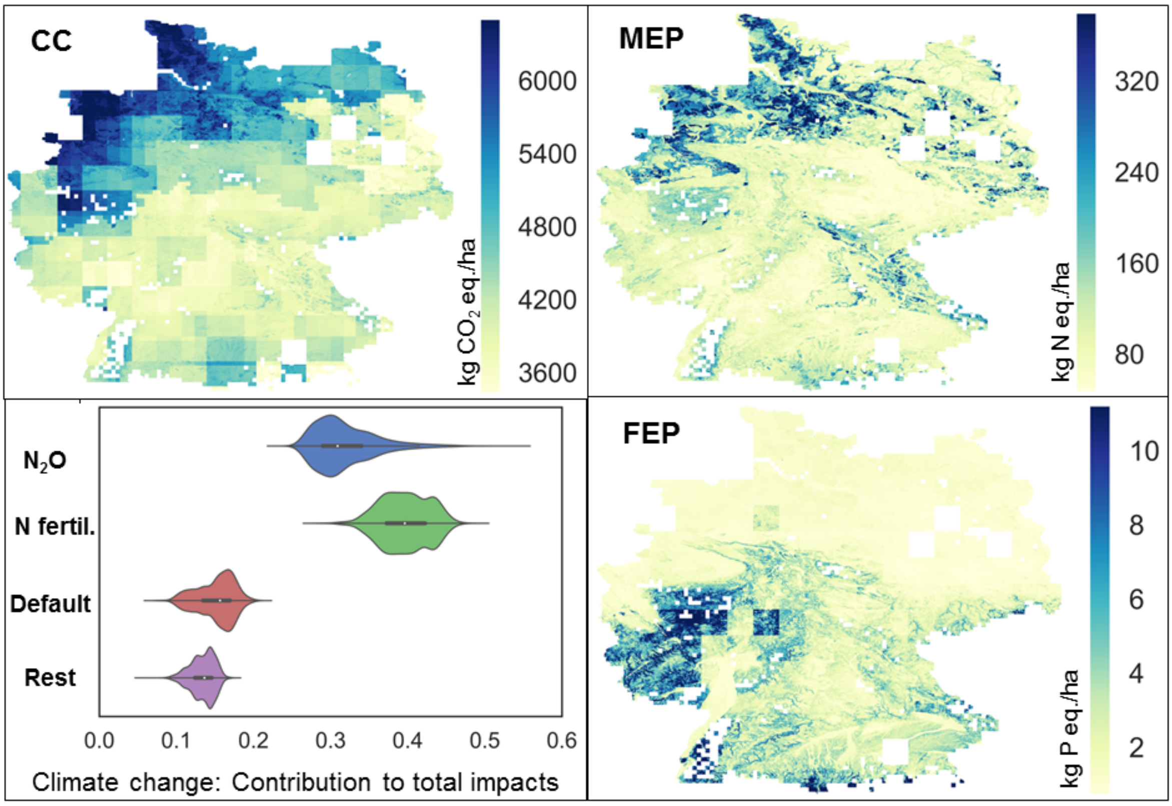


Figure 4-2: Heat map showing the spatially explicit environmental impacts (in kg impact per hectare) of roughly 588’000 UPDs for the LCIA indicators CC, MEP and FEP. White areas represent blind spots caused by data gaps. The violin plot (left bottom) shows the relative contribution of four types of exchanges to the total climate change impact: the two exchanges contributing the most (N_2O and N-based fertilizers), all exchanges modeled with default data (Default) and all remaining exchanges (Rest). The vertical extent of the “violins” shows the frequency of data points.

[58] Roughly 6% of the UPDs (or 40’000 UPDs) were excluded due to data gaps. The extrapolation of missing data was not in the scope of this article.

CC impacts are dominated by N_2O emission (resulting from N-based fertilizer application) and the energy intensive production of N-based mineral fertilizer; both typically cause more than 70% of the impacts (see violin plot). Consequently, the spatial distribution of CC impacts correlates largely with the application intensity of N-based mineral fertilizer. The impacts according to MEP and FEB are dominated by the contributions of nitrate (90%) and phosphorus emissions (50%), respectively. Nitrate emissions are dependent on many factors. Therefore, the MEP heat map indicates an impact pattern much more diverse than the pattern observed for CC impacts. Phosphorus emissions are largely dependent on soil erosion, which is particularly large for the federal state of Rhineland-Palatinate.

Table 4-3 contrasts the mean of important inventory flows with corresponding values in existing rapeseed UPDs and literature data. In general, our results appear reasonable; though often higher than the values from comparable inventories.

Table 4-3: Comparison with generic inventory data from different LCI databases.

Exchange flow	Unit	Nem	AB	AF	EI3.2	Our result	
						Ø	CV
Nitrogen fertilizer	kg N/ha		162	200	113	207	0.22
Phosphate fertilizer	kg P_2O_5/ha		34	46	54	66	0.32
Potassium chloride	kg K_2O/ha		25	82	45	73	0.19
Nitrate (NO_3)	kg NO_3-N/ha	30 – 140	32	71	12	110	0.73
N_2O	kg N_2O-N/ha	0.5-2.5*	2.75	3.12	2.72	3.4	0.31
Phosphorus, river	kg / ha		0.53	n.a.	0.12	0.77	1.47

Nem = literature review of (Nemecek et al., 2014); AB = Agribalyse, rapeseed production in France; AF=AgriFootprint, rapeseed production in Germany; EI3.2 = ecoinvent version 3.2, rapeseed production in Germany; CV = coefficient of variation. *Asterisks indicate that values are not rapeseed specific.

The coefficient of variation (CV) indicates a high spatial sensitivity of Phosphorus and Nitrate emissions, i.e., the size of these flows varies greatly throughout space. This confirms that the spatially explicit computation of these flows is important to obtain accurate cultivation process datasets.

4.4 Discussion

The case study presented above is probably the most comprehensive cradle-to-gate LCA on climate change and eutrophication impact of rapeseed cultivation in Germany. However, it still relies on many assumptions such as the use of standard values for crop tillage and management practice to compute phosphorous flows or the static modeling of field operations on the basis of default data obtained from the original rapeseed cultivation datasets in the ecoinvent database.

While the use of default values in the case study decreases the degree of regionalization, the general possibility to revert to such defaults should be considered as one of the main strength of the framework. It facilitates the fine-tuning of the degree of regionalization to the spatial sensitivity of exchange flows and to the availability of spatial models and input data. On the other hand, computations in "high resolution" (30 arc seconds and lower) for large geographic areas such like the US or Brazil can currently not be performed on a modern laptop computer as they are to time intensive [59]. However, computations in such resolutions are important for the appropriate consideration of certain spatial characteristics such like the length slope factors and erodibility. We are currently implementing a more efficient approach for the merging of the raster data into one parameter table (step 2)—the most time intensive step—which is expected to facilitate computations in "high resolutions" for large geographies also on a modern laptop computer. In addition, the relevance of the spatial repository should be updated. For example, the spatial data on harvested rapeseed area (see Table 1, no. 6) referring to the year 2000 does not consider the recent 25% increase to 1.5 million hectare (FAOSTAT, 2016). Finally, the framework doesn't support UPD export in common LCA data formats such like the ILCD or the ecospold2 format.

Despite these limitations, the overall framework harbors great potential for improving LCI modeling, as the following use cases may show. First, the framework might be relevant in every project setting where the assessment of the environmental impacts of spatially sensitive processes are key, i.e., when food, feed, fibre and (bio-)fuel production are under study. In this context, it offers an improved basis for decision making as it allows to explicitly consider and compute variations in micro-spatial conditions, which are typically ignored in the common "unit world" paradigm of LCA. These variations may be relevant, for example, when a large-scale bioenergy producer wants to compare the environmental performance of alternative bioenergy-cropping systems in an explicit spatial setting.

Second, the framework can be used as a tool to improve LCI databases by adding more accurate, average agricultural UPDs. To date, UPDs in such databases are mostly site-generic. The consistent, spatially explicit bottom-up computation of UPDs allows for the generation of more accurate

[59] For Brazil, we have to merge about roughly 25 x 20 million data entries. With the current merging approach, this would require more than a month on a modern laptop computer.

averages for relatively homogeneous regions. That is, the spatial scale (i.e. the region covered) of an average UPD does not necessarily have to follow political (country or state) boundaries but could be determined by a spatial cluster analysis which considers location and the spatial variation in exchange flow range. As a welcome side-effect, the framework facilitates a better evaluation of spatial uncertainty of exchange flows[60] as it allows the exploration of the spatial sensitivity of each regionalized exchange flow for any spatial extent. Improving both the representativeness of and spatial uncertainty information in agricultural UPDs is an improvement at the very foundation of LCA, as it makes the entire background system involved in every LCA more reliable.

Third, our framework offers a basis for the integration of more sophisticated emission models that were developed outside the LCA context. Emission models applied in agricultural UPDs are often adjusted to the "averaging" nature of LCA, i.e., they lack important compartments due to their generic spatial and temporal orientation. For example, the computation of nitrate ignores losses with waterborne and windborne sediment. However, as shown in the result section, the emissions in agricultural UPDs often contribute significantly to the total environmental impact. Our framework offers a modular and expandable test bed for the application advanced models. This could improve the representativeness and accuracy of the emissions recorded in agricultural UPDs.

Fourth, the framework could be used as a background computation engine in web-based LCA tools (often called "footprinters"). For example, our framework could improve the usability and accuracy of the RSB tool (Reinhard et al., 2011). The RSB tool is a web-based tool for the carbon footprint computation of various user-defined biofuel pathways. The user has to enter a great amount of context-specific information (soil type, land use, precipitation, etc.) into a questionnaire. Our framework could reduce these data entries and improve the accuracy of inventory computations.

Finally, our framework delivers regionalized UPDs, which is the fundament also for regionalized LCIA.

4.5 Conclusion

We presented a framework for the automated generation of site-specific agricultural unit process datasets (UPDs). Our approach is to transform publicly available, spatial raster data into site-

[60] To date, exchange flow specific variability in LCI databases (used for Monte-Carlo analysis) is largely based on rough estimates.

specific UPDs on the basis of the WFLDB Guidelines. Although the framework has been demonstrated by means of a case study on rapeseed cultivation in Germany, it can be applied to any major crop or country on the globe.

Acknowledgments

We thank our colleague Mireille Faist-Emmenegger for her support with WFLDB emission models and the three anonymous reviewers for their helpful comments.

4.6 References

Carlson, R., Tillman, A.-M., Steen, B., Löfgren, G.: Lci data modelling and a database design. Int. J. Life Cycle Assess. 3, 106–113 (1998).

Dresen, B., Jandewerth, M.: Integration of spatial analyses into LCA—calculating GHG emissions with geoinformation systems. Int. J. Life Cycle Assess. 17, 1094–1103 (2012).

FAOSTAT: http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor.

Finnveden, G., Hauschild, M.Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., Suh, S.: Recent developments in Life Cycle Assessment. J. Environ. Manage. 91, 1–21 (2009).

Geyer, R., Lindner, J.P., Stoms, D.M., Davis, F.W., Wittstock, B.: Coupling GIS and LCA for biodiversity assessments of land use. Int. J. Life Cycle Assess. 15, 692–703 (2010)(a).

Geyer, R., Stoms, D., Lindner, J., Davis, F., Wittstock, B.: Coupling GIS and LCA for biodiversity assessments of land use. Int. J. Life Cycle Assess. 15, 454–467 (2010)(b).

Guinée, J.B., Gorrée, M., Heijungs, R., Huppes, G., Kleijn, R., de Koning, A., van Oers, L., Sleeswijk, A.W., Suh, S., de Haes, H.A.U., de Bruijn, H., van Duin, R., Huijbregts, M.A.J.: Handbook on Life Cycle Assessment. Operational Guide to the ISO Standards. (2002).

Hauschild, M.: Spatial Differentiation in Life Cycle Impact Assessment: A decade of method development to increase the environmental realism of LCIA. Int. J. Life Cycle Assess. 11, 11–13 (2006).

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R.: SoilGrids1km — Global Soil Information Based on Automated Mapping. PLoS ONE. 9, e105992 (2014).

IIASSA/FAO: Global Agro-ecological Zones (GAEZ v3.0). IIASA, Laxenburg / FAO, Rome (2012).

Monfreda, C., Ramankutty, N., Foley, J.A.: Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Glob. Biogeochem. Cycles. 22, (2008).

Mueller, N.D., Gerber, J.S., Johnston, M., Ray, D.K., Ramankutty, N., Foley, J.A.: Closing yield gaps through nutrient and water management. Nature. 490, 254–257 (2012).

Mutel, C.: A new open source framework for advanced life cycle assessment calculations, https://brightwaylca.org/.

Mutel, C.L., Hellweg, S.: Regionalized Life Cycle Assessment: Computational Methodology and Application to Inventory Databases. Environ. Sci. Technol. 43, 5797–5803 (2009).

Mutel, C.L., Pfister, S., Hellweg, S.: GIS-based regionalized life cycle assessment: how big is small enough? Methodology and case study of electricity generation. Environ. Sci. Technol. 46, 1096–1103 (2012).

Natural Earth: Admin 0 - Country data, ww.naturalearthdata.com/http//www.naturalearthdata.com/download/110m/cultural/ne_110m_admin_0_countries_lakes.zip.

Nemecek, T., Bengoa, X., Lansche, J., Mouron, P., Riedener, E., Rossi, V., Humbert, S.: Methodological Guidelines for the Life Cycle Inventory of Agricultural Products. World Food LCA Database (WFLDB), Version 3.0, http://www.quantis-intl.com/microsites/wfldb/files/WFLDB_MethodologicalGuidelines_v3.0.pdf, (2015).

Nemecek, T., Schnetzer, J., Reinhard, J.: Updated and harmonised greenhouse gas emissions for crop inventories. Int. J. Life Cycle Assess. 1–18 (2014).

Panagos, P., Borelli, P., Meusburger, K.: A New European Slope Length and Steepness Factor (LS-Factor) for Modeling Soil Erosion by Water. Geosciences. 5, 117–126 (2015).

Panagos, P., Meusburger, K., Ballabio, C., Borrelli, P., Alewell, C.: Soil erodibility in Europe: A high-resolution dataset based on LUCAS. Sci. Total Environ. 479–480, 189–200 (2014).

Pfister, S., Bayer, P., Koehler, A., Hellweg, S.: Environmental Impacts of Water Use in Global Crop Production: Hotspots and Trade-Offs with Land Use. Environ. Sci. Technol. 45, 5761–5768 (2011).

Potter, P., Ramankutty, N., Bennett, E.M., Donner, S.D.: Characterizing the Spatial Patterns of Global Fertilizer Application and Manure Production. Earth Interact. 14, 1–22 (2010).

Potter, P., Ramankutty, N., Bennett, E.M., Donner, S.D.: Global Fertilizer and Manure, Version 1: Nitrogen in Manure Production, http://dx.doi.org/10.7927/H4KH0K81, (2011)(a).

Potter, P., Ramankutty, N., Bennett, E.M., Donner, S.D.: Global Fertilizer and Manure, Version 1: Phosphorus in Manure Production, http://dx.doi.org/10.7927/H49Z92TD, (2011)(b).

Reinhard, J., Emmenegger, M.F., Widok, A.H., Wohlgemuth, V.: RSB tool: a LCA tool for the assessment of biofuels sustainability. In: Proceedings of the Winter Simulation Conference. pp. 1048–1059. Winter Simulation Conference, Phoenix, Arizona (2011).

Scherer, L., Pfister, S.: Modelling spatially explicit impacts from phosphorus emissions in agriculture. Int. J. Life Cycle Assess. 20, 785–795 (2015).

Seto, K.C., Reenberg, A., Boone, C.G., Fragkias, M., Haase, D., Langanke, T., Marcotullio, P., Munroe, D.K., Olah, B., Simon, D.: Urban land teleconnections and sustainability. Proc. Natl. Acad. Sci. (2012).

YANG, X.. P., W.M...THORNTON, P.E...JAIN, A.K..: Global Gridded Soil Phosphorus Distribution Maps at 0.5-degree Resolution, http://dx.doi.org/10.3334/ORNLDAAC/1223, (2014).

Zah, R., Gmuender, S., Gauch, M.: Evaluación del ciclo de vida de la cadena de producción de biocombustibles en Colombia. Banco Interamericano de Desarrollo (BID) (2012).

PART III – CONCLUSION

5 CONCLUSION

Unit process datasets form the foundation of every LCA application. However, the ability to generate unit process datasets is constrained by the high cost of primary data collection. Despite more than 15 years of common data collection efforts, the base data for LCA is still incomplete and uncertain. The ecoinvent database contains 13,000 datasets—at first sight a large number, but in fact no match for the complex economic system that LCA strives to model from the bottom up, even assuming a high level of abstraction.

This dissertation develops two novel IS-based solutions for organizing efforts around data collection in LCA more efficiently and effectively: (i) a statistical prioritization approach that directs LCI database improvement efforts toward datasets of key importance in terms of their potential influence on overall database quality, and (ii) a computational framework that facilitates the automated generation of regionalized cultivation datasets on the basis of publicly available spatial (raster) data.

With regard to the statistical prioritization, our argument is as follows. Considering the high cost of collecting primary data, knowing exactly which processes are most important is a prerequisite for effectively improving LCI databases. We develop a statistical prioritization framework that helps organizations maintaining LCI databases to identify the key processes to be prioritized for improvement. Knowing the relative importance of the different processes in a LCI database makes it possible to act in a targeted fashion, improving those data elements with the greatest influence on overall database quality. For this reason, our first research question is:

RQ1: How can we identify and use relative process importance in LCI databases to organize data collection efforts more effectively?

We address this research question as follows:

- **A novel approach for database-wide contribution analysis (CA)**. Although some studies use CA to analyze certain features of LCI databases (Frischknecht et al. 2007; Lesage and Samson 2013; Rørbech et al. 2014), this has not been formalized from the point of view of prioritizing LCI database improvements. Our method computes the relative contribution of each unit process according to two perspectives: a *causer* perspective and a *connector* perspective. These two perspectives represent complementary tools that we can use to achieve our goal of identifying important processes throughout the database. In the causer perspective, we focus exclusively on the causative elements of each unit process, that is, the direct exchanges with the environment (resources consumptions and emissions) that cause environmental impacts. This perspective helps pinpoint those processes with consistently large contributions in terms of environmental interventions. In the connector perspective, we focus exclusively on the connecting elements of each unit process, that is, the intermediate flows from other processes within the technosphere. This perspective helps us pinpoint those unit processes that are consistently linked to large upstream contributions. We adapt and expand the LCA standard matrix inversion approach so that they include these two perspectives. For the causer perspective, the key adaptations are minor modifications that facilitate an in-depth analysis of the contributions per process and the simultaneous computation of the entire database. For the connector perspective, we develop a new set of matrix equations.

- **A novel set of summary measures for investigating relative process importance.** Summary measures "condense a certain aspect of a large set of numbers into one or perhaps a few items" (Heijungs and Suh 2002). Using the database-wide CA as a starting point, we develop a set of summary metrics that facilitates the identification of key datasets and summarizes the inequalities in process importance. We use a mirror image of the *Lorenz* curve (MLC)—the most popular tool for comparing income and wealth inequality in the field of economics—and several related measures of inequality, such as the Gini coefficient and the concentration ratio, to assess inequality and concentration in relative process importance. The points on the MLC represent statements such as "the top 1% of the processes in the database already account for 85% of the total contribution". The corresponding data structure offers a powerful way to analyze relative process importance and prioritize across the relative process importance for many LCIA indicators. To the best of our knowledge, the MLC and associated measures of inequality have not previously been applied in the realm of LCA.

- **A novel ranking algorithm to determine dataset importance across different LCIA indicators**. Relative process importance is a function of the applied LCIA indicator. In other words, different LCIA indicators will prioritize different processes: Prioritizing for just one LCIA indicator would lead to a one-sided improvement of the LCI database. A robust prioritization must therefore consider many LCIA indicators. Our method makes it easier to calculate an overall rank of unit processes considering their importance across any given set of LCIA indicators. Using the MLC and an arbitrary set of thresholds for the total contribution—for example, 10%, 20%, 30%, and so forth—the algorithm identifies the minimum amount of processes required to exceed a given contribution for each LCIA indicator. We determine the importance of a particular unit process first by its size class (threshold) in the MLC (for example, is it required to exceed the first or the second threshold?), and only then by its LCIA support (how many LCIA indicators point to the process of interest?). This procedure guarantees that each LCIA indicator is treated equally, as it ensures that all the processes required to exceed a given threshold are considered. We have not been able to identify a similar approach for ranking data elements in the literature. While this method was developed in the context of LCI database prioritization, it could be used to support the interpretation phase of any standard LCA application.

- **A demonstration and evaluation of the new prioritization method in a proof-of-concept application and a comprehensive analysis of a LCI database.** We demonstrate and evaluate the validity, performance, and utility of the above-mentioned contributions in the form of a proof-of-concept application and a comprehensive analysis of Version 3.1 of the ecoinvent database. Both tests validate the prioritization framework, providing evidence for the fact that a tiny, robust set of unit processes is responsible for most of the contribution across the entire LCI database, and even across the broad spectrum of all modern LCIA indicators (Article I, Section 2.4.5). Focusing research efforts on increasing information density in these processes, which are of systemic importance, offers a fresh starting point for improving the entire database systematically and effectively. The ranking of most important processes we present (SI2) strengthens the basis for decision-making used by those managing the data, who can systematically process the list depending on the time and resources available, and decide for each unit process if and how it should be improved (Article II, SI2).

Overall, our research provides useful new design knowledge in the form of operational principles that can be adapted and applied in as yet unstudied fields (see Section 1.6.1). It focuses attention on the significance of LCI database analysis and shows that detailed insights into relative process importance in existing LCI databases provide a valid foundation for more effective data collection. The LCA community can benefit greatly from this insight. The administration teams behind all LCI databases can organize their data collection strategy more effectively with the help of our prioritization framework, thereby improving the quality of the key datasets that build the very foundation of practically every LCA application today.

With regard to the regionalization framework, our argument is as follows. The high cost of data collection limits the representativeness and accuracy of LCA, particularly in the field of agriculture, where exchange flows are highly sensitive to the natural variability of the local environment. We develop a regionalization framework that facilitates the automatic generation of regionalized unit process datasets using publicly available spatial data and background data from the ecoinvent database. While recent studies (Geyer et al. 2010b, a; Reinhard et al. 2011; Zah et al. 2012; Dresen and Jandewerth 2012; Scherer and Pfister 2015) have already succeeded in the case-specific parameterization of the spatial properties of selected exchanges, we are not aware of any research that provides a general framework for regionalized LCI modeling generating comprehensive, regionalized unit process datasets. For this reason, our second research question is:

RQ2: How can we automatically process publicly available spatial data into regionalized comprehensive unit process datasets that consider existing background data from LCI databases?

We address this research question as follows:

- **A novel regionalization framework for automatically generating regionalized cultivation datasets using publicly available spatial (raster) data**. The method processes publicly available spatial (raster) data into comprehensive unit process datasets using background data from the ecoinvent database and the emission models from the World Food Life Cycle Database (WFLDB) guidelines. This yields comprehensive unit process datasets primarily thanks to two improvements: the capacity to process different raster data resolutions (30x30 arc second, 5x5 min, and so on) and formats (Geotiff, NetCDF, and so on) and the connection to the ecoinvent database. We develop functions for processing different spatial data formats into Pandas data frames—a Python Data Analysis Library. We can merge different data frames on the basis of a shared index consisting of latitude-longitude combinations. Using this index as a basis, we can efficiently join data frames of different resolutions into one and the same data frame. We process and expand this data into a comprehensive regionalized unit process dataset using background data from the ecoinvent database. The integration of ecoinvent data makes it possible to consider the entire background life cycle (field operations, production of fertilizers, and so on). It also facilitates the generation of complete unit process datasets, even when detailed data is lacking or cannot be computed.
- **A demonstration and evaluation of the regionalization framework in a case study**. We apply the framework in a case study on rape seed cultivation in Germany in order to

explore its utility in a practical setting. With 580,000 datasets, the case study presented is likely the most comprehensive cradle-to-gate LCA on climate change and eutrophication impact of rapeseed cultivation in Germany. Highly important, spatially-sensitive exchange flows such as N_2O, nitrate, and phosphorous are computed explicitly for each grid cell (approximately 1x1 km) using a wide array of spatial-specific input parameters, such as precipitation, soil organic carbon content, erodibility, length-slope factor, and the like. This makes it easier to consider spatial conditions that are not generally accounted for when agricultural datasets are generated, typically focusing on country averages. The case study shows that regionalization in the field of agriculture matters: Contribution to climate change varies by a factor of two, while marine and freshwater eutrophication varies by a factor of five. To a large extent this variation is related to the variability of nitrate, phosphorous, and N_2O emissions. This confirms that the spatially explicit computation of these flows is important for obtaining accurate cultivation process datasets.

Overall, our research produces beneficial new design knowledge that provides a starting point for increasing the utility of LCA, particularly in the context of other environmental system analysis tools. Our framework enhances the efficiency and quality of unit process dataset generation in the area of agricultural LCA by facilitating the automated generation of high-resolution agricultural process datasets for all major crops and all regions of the world. It improves the geographical representativeness, completeness, and reproducibility of agricultural datasets and offers new possibilities for aggregating and analyzing them. It focuses attention on the significance of spatial data in the area of LCI modeling and may be relevant for any project where assessing the environmental impact of spatially sensitive processes is key, in other words, where the object of study is food, feed, fiber, and (bio)fuel production. In this context, it offers an improved basis for decision-making as it makes it possible to explicitly consider and compute variations in micro-spatial conditions, typically ignored in the common "unit world" paradigm of LCA.

LCA is an ambitious field of study, one that requires the collection and processing of a vast amount of data and has to deal with the enormous complexity of the socio-technical system (Davis et al. 2010). To realize its full potential, a more deliberate strategy for generating and curating data is needed. The research presented here shows that both prioritization and regionalization are valid, useful approaches for improving the foundation (including the data basis) of decision-making based on LCA. While this represents just a small direct contribution to the overall quality of LCA, we believe that it demonstrates the immense possibilities and potential offered by the smart use of computational approaches in LCA.

5.1 References

Dresen, Boris, and Michael Jandewerth. 2012. "Integration of Spatial Analyses into LCA—calculating GHG Emissions with Geoinformation Systems." The International Journal of Life Cycle Assessment 17 (9): 1094–1103. doi:10.1007/s11367-011-0378-3.

Frischknecht, Rolf, Hans-Jörg Althaus, Christian Bauer, Gabor Doka, Thomas Heck, Niels Jungbluth, Daniel Kellenberger, and Thomas Nemecek. 2007. "The Environmental Relevance of Capital Goods in Life Cycle Assessments of Products and Services." International Journal of Life Cycle Assessment 12 (1): 7–17.

Geyer, Roland, Jan P. Lindner, David M. Stoms, Frank W. Davis, and Bastian Wittstock. 2010. "Cou-pling GIS and LCA for Biodiversity Assessments of Land Use." The International Journal of Life Cycle Assessment 15 (7): 692–703. doi:10.1007/s11367-010-0199-9.

Geyer, Roland, DavidM. Stoms, JanP. Lindner, FrankW. Davis, and Bastian Wittstock. 2010. "Coupling GIS and LCA for Biodiversity Assessments of Land Use." The International Journal of Life Cycle Assessment 15 (5): 454–67. doi:10.1007/s11367-010-0170-9.

Heijungs, R., and Sangwon Suh. 2002. The Computational Structure of Life Cycle Assessment. Kluwer Academic Publishers, Dordrecht.

Lesage, Pascal, and Réjean Samson. 2013. "The Quebec Life Cycle Inventory Database Project." The International Journal of Life Cycle Assessment, 1–8. doi:10.1007/s11367-013-0593-1.

Reinhard, Juergen, Mireille Faist Emmenegger, Andi H. Widok, and Volker Wohlgemuth. 2011. "RSB Tool: A LCA Tool for the Assessment of Biofuels Sustainability." In Proceedings of the Winter Simulation Conference, 1048–59. Phoenix, Arizona: Winter Simulation Conference.

Rørbech, Jakob T., Carl Vadenbo, Stefanie Hellweg, and Thomas F. Astrup. 2014. "Impact Assessment of Abiotic Resources in LCA: Quantitative Comparison of Selected Characterization Models." Environmental Science & Technology 48 (19): 11072–81. doi:10.1021/es5023976.

Scherer, Laura, and Stephan Pfister. 2015. "Modelling Spatially Explicit Impacts from Phosphorus Emissions in Agriculture." The International Journal of Life Cycle Assessment 20 (6): 785–95. doi:10.1007/s11367-015-0880-0.

Zah, Rainer, Simon Gmuender, and M. Gauch. 2012. "Evaluación del ciclo de vida de la cadena de producción de biocombustibles en Colombia." Estrategias de energía sostenible y biocombus-tibles para Colombia ATN/JC-10826-CO y ATN/JF-10827-CO. Banco Interamericano de Desarrollo (BID).

PART IV – APPENDIX

6 GLOSSARY

LCI database	"An LCI database is a system intended to organize, store, and retrieve large amounts of digital LCI datasets easily. It consists of an organized collection of LCI datasets that completely or partially conform to a common set of criteria including methodology, format, review, and nomenclature. The database will allow for interconnection of individual datasets to create LCI models. The computed results can be used with identified life cycle impact assessment (LCIA) methods for life cycle assessment (LCA). Databases are managed using database management systems, which store database contents, allowing data creation and maintenance, search, and other access" (Sonnemann and Vigon, 2011, p. 86). An LCI database is distinguished from a dataset library.
Dataset library	"A collection of datasets that may not conform to common criteria and do not allow for interconnections and common applications for LCA or LCIA purposes. An example of a dataset library is the United Nations Environment Programme/Society of Environmental Toxicology and Chemistry (UNEP/SETAC) Database Registry" (Sonnemann and Vigon, 2011, p. 86) .
Unit process	"Smallest element considered in the life cycle inventory analysis for which input and output data are quantified" (ISO, 2006, chap. 3.34). "A unit process dataset is obtained as a result of quantifying inputs and outputs in relation to a quantitative reference flow from a process. These inputs and outputs are generated from mathematical relationships that operate on raw data that have not previously been related to the same reference flow" (Sonnemann and Vigon, 2011, p. 54).
Product system	"Collection of unit processes with elementary and intermediate flows, performing one or more defined functions, and which models the life cycle of a product"(ISO, 2006, chap. 3.28).
Reference flow	"Measure of the outputs from processes in a given product system required to fulfil the function expressed by the functional unit" (ISO, 2006, chap. 3.29).
Functional unit	"Quantified performance of a product system for use as reference unit" (ISO, 2006, chap. 3.20).
Spatial scale	In this dissertation we use spatial scale to refer to the extend or size of an area for which a particular agricultural unit process dataset represent a valid representation or aggregation.
Instantiations	"The physical realizations that act on the natural world, such as an information system that stores, retrieves, and analyzes customer relationship data." (Gregor and Hevner, 2013, p. A3) .
Methods	"Algorithms, practices, and recipes for performing a task." (Gregor and Hevner, 2013, p. A3) .
Constructs	"Provide the vocabulary and symbols used to define and understand problems and solutions"(Gregor and Hevner, 2013, p. A3) .
Models	"Designed representations of the problem and possible solutions" (Gregor and Hevner, 2013, p. A3) .
Precision	"Measure of the variability of the data values for each data expressed (e.g. variance)" (ISO, 2006, chap. 4.2.3.6.2 c)). This definition relates to the statistical meaning of stochastic uncertainty, i.e., the degree to which further measurements or calculations done by different experts show the same results" (EC-JRC, 2011, p. 325). A low variance indicates a high precision.
Appropriateness	"How far a data set in a system model represents a truly required process or product"(EC-JRC, 2011, p. 325)
Representativeness	"How far the dataset is depicting the functional unit(s) and/or reference flows(s) of the process" (EC-JRC, 2011, p. 326). Representativeness of a unit process dataset is determined by its technological, time-related and geographical representativeness (EC-JRC, 2011)
Technological	„Degree to which the data set reflects the true population of interest regarding

representativeness	technology (…)“ (EC-JRC, 2011, p. 329).
Geographical representativeness	„Degree to which the data set reflects the true population of interest regarding geography(…)“(EC-JRC, 2011, p. 329).
Time-related representativeness	„Degree to which the data set reflects the true population of interest regarding time / age of the data“ (EC-JRC, 2011, p. 329).
Accuracy	“Refers to the degree of closeness of a measured or calculated quantity to its actual (true) value. Accuracy hence captures the technological, geographical and time-related representativeness as well as appropriateness and consistency of methods and their use” (EC-JRC, 2011, p. 325).
Data quality aspects	Accuracy, precision and completeness (EC-JRC, 2011, p. 325).
Completeness	“Percentage of flow that is measured or estimated” (ISO, 2006, chap. 4.2.3.6.2 e))

References

EC-JRC, 2011. International Reference Life Cycle Data System (ILCD) Handbook- Recommendations for Life Cycle Impact Assessment in the European context. First edition November 2011. EUR 24571 EN. Luxemburg. Publications Office of the European Union;

Gregor, S., Hevner, A.R., 2013. Positioning and Presenting Design Science Research for Maximum Impact. MIS Q. 37, 337–355.

ISO, 2006. ISO 14044. Environmental management - Life cylce assessment - Requirements and guidelines.

Sonnemann, G., Vigon, B. (eds)., 2011. Global Guidance Principles for Life Cycle Assessment Data-bases: A Basis for Greener Processes and Products (Shonan Guidance Principles).

7 LIST OF FIGURES

8 List of Tables

9 Curriculum Vitae

Personal Information

Name:	Jürgen Reinhard
Nationality:	German
Date of birth:	25.03.1980
E-Mail:	juergen.reinhard@uzh.ch

Awards

Sept. 16	Best Paper Award: EnviroInfo2016 for *"Regionalized LCI modeling: A Framework for the integration of spatial data into Life Cycle Assessment"*
Nov. 09	Tiburtius-Award: Highest award for excellent Master's thesis at Universities in Berlin in 2008.

Education

Oct. 12 – ongoing	Ph.D. in Information Systems, Informatics and Sustainability Research, University of Zurich / Empa.
Oct. 10 – Sept. 12	Master of Science (fast track) in Information Systems: University of Zurich.
Oct. 05 – March 08	Master of Science in Industrial Environmental Computing: University of Applied Science (HTW) Berlin.
Oct. 02 – Oct. 05	Bachelor of Science in Industrial Environmental Computing: University of Applied Science (HTW) Berlin.

Professional Work Experience (extract)

July 13 – ongoing	Quantis – Switzerland/Germany: • Project manager | Conduct various environmental product declarations in the domain of sanitary appliance
Nov. 11 – Dec. 16	Swiss Centre of Life Cycle Inventories (ecoinvent): • LCI expert | Maintaining and developing the ecoinvent data quality guidelines and monitoring their implementation. • Editor | Review, validate and edit new datasets.
May 08 – June 13	Empa - Life Cycle Assessment (LCA) and Modeling Group: • LCI specialist | Development of a land use change methodology for the generation of harmonized crop inventories. • Software designer | Development of a web-based tool for the assessment of biofuels sustainability. • LCA expert, ISO/TC 248 (Sustainability criteria for bioenergy) | Head of delegation, CH